Katrin Hänsel

The potential of emerging wearable physiological sensing in the space of human-subject studies

submitted as

Thesis - Doctor of Philosophy

at

Queen Mary University of London

School of Electronic Engineering and Computer Science

Bibliographic Description

Hänsel, Katrin:
The potential of emerging wearable physiological sensing in the space of human-subject studies,
2019 – London, Queen Mary University of London, School of Electronic Engineering and Computer Science, Thesis - Doctor of Philosophy

Statement of Originality

I, Katrin Hänsel, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Katrin Hänsel Date: 14 June 2019

Abstract

In recent years, novel sensing means in the form of smartwatches and fitness trackers with integrated sophisticated sensing emerged on the consumer market. While their primary purpose is to provide consumers with an overview of rough-grained health-related metrics, these signals offer to pick up fine-grained changes within the human body. This thesis considers the suitability of these novel wearable sensing devices to be used in affective research.

Firstly, and based on the work with concrete state-of-the-art wearables, issues around the access of research-suitable data are discussed. The findings are put in context by examining common wearable device architectures and data access means provided. The discussion concludes with aspects researchers need to consider when seeking data access from state-of-the-art or future wearables.

Secondly, two research probes explore the application of four exemplary devices to detect stress and affect in the wild and in the lab. Issues around the data reliability and participant comfort arose. The experiences are reflected upon to provide researchers with a summary of aspects to consider when applying wearable sensing devices in affective research.

Lastly, this thesis contributes a Design Space for Physiological Measurement Tools. This design space was evaluated with a qualitative study enquiring research experts experiences. The resulting Design Space presents seven distinct dimensions of factors to consider when choosing a wearable sensing device for research. This design space has been applied to a novel sensing device which was used for a study on interpersonal synchrony.

The insights and the 'Design Space for Physiological Measurement Tools' provide researchers with a tool to apply when they consider to use wearable physiological sensing devices in research.

Table of Contents

V Application of the Design Space in an Exemplary Setting

VI Conclusion and Future Work

VII Appendix and Bibliography

List of Figures

List of Tables

Glossary and Abbreviations

accelerometer are sensors to measure acceleration (i.e., the rate of change in velocity of an object). Most novel smartphones and wearables are equipped with accelerometer sensors.

ANS *(Autonomous Nervous System)* regulates unconscious, bodily functions like organs, hormone responses, etc.

API *(Application Programming Interface)* describes the library endpoints.

app is the short term for applications; and it refers in particular to mobile phone applications.

App Store is a distribution platform for consumers to purchase and download apps. In terms of this report, we refer specifically to the Apple App Store.

Apple Watch is a wrist-worn smartwatch from Apple (`www.apple.com`). It is one of the devices evaluated in Chapters 4 and 5.

arousal psychological arousal refers to the state of alertness of a user; a low arousal can be related to tiredness/boredom while a high arousal level relates to excitement/stress.

AWSense is an *Apple Watch* sensing framework by the thesis author Hänsel et al. (2017).

BCI *(Brain-Computer Interface)* is a communication interface between the human brain and an *Electroencephalogram (EEG)* sensing unit interpreting the signal.

BFI-10 *(10-item Big Five Inventory)* is a short scale to assess the *Big Five Personality traits* developed by Rammstedt and John (2007).

Big Five Personality traits are a common taxonomy for personality in 5-dimensions (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism).

BioTrace+ is a software provided by MindMedia (2017). It provides functionalities to configure, record, real-time visualise and export the collected sensing data from the *Nexus 10 MK2* biofeedback device.

BLE *(Bluetooth Low Energy)* is a wireless personal area networking technology for short-range communication by the Bluetooth Special Interest Group (SIG) (Bluetooth Special Interest Group, 2010).

Circumplex Model of Affect by Russell (1980) classifies affective states in two dimensions: pleasure and arousal. A detailed description can be found in Section 2.3.1.

CrowdSense is an iOS mobile sensing tool based on SensingKit by Katevas et al. (2014).

CSV *(Comma Separated Values)* is a file format where data is stored in a human-readable form separated by commas.

DIKW *(Data–Information–Knowledge–Wisdom)*, also called *Wisdom Hierarchy* or *Knowledge Pyramid*, is a framework for hierarchical information and knowledge modelling originated by Ackoff (1989) and extensively discussed by Rowley (2007). The pyramid-shaped hierarchy is shaped based on the nature of information and knowledge with an increasing degree of contextualisation and interpretation. The model has been applied to wearable sensing data in Section 4.1.

DV *(dependent variable)* is the measured variable within an experiment.

E4 wristband is a wrist-worn sensing device with focus on research applications by Empatica Inc.. It features *accelerometer*, *Photoplethysmography (PPG)* heart rate, *Electrodermal Activity (EDA)*, and skin temperature sensors. Within this thesis, it is used in the study of 7.

ECG *(Electrocardiogram)* picks up the electrical activity of the heart.

ECG *(Electroencephalogram)* picks up the electrical activity of the heart.

EDA *(Electrodermal Activity)* describes the changed in the electrical properties of the skin.

EMG *(Electromyogram)* picks up the electrical activity of muscles.

EmoRate is an iOS and Apple Watch *app* developed in this project; the corresponding study is presented in Section 5.1.

EQ Scale *(Empathy Quotient Scale)* is a psychological self-report measure to assess empathy regarding an individual (Baron-Cohen and Wheelwright, 2004). Different versions of varying length exist. A short form with 8-items by Loewen et al. (2010) has been used in study of 7. first.

eSIM *(embedded SIM)* is a programmable SIM card embedded in a mobile or wearable device to allow to authenticate subscribers to a mobile network.

ESM *(Experience Sampling Method)* by Larson and Csikszentmihalyi (1983) describes a research method where participants are asked to note their right-now state. This can comprise emotions, the current situation or their experiences at the very moment.

exergame is a game to facilitate exercise and movement by mostly using it as a game element.

GATT *(Generic Attributes)* is a generic data structure that is exposed to connected *Bluetooth Low Energy (BLE)* devices Bluetooth Special Interest Group (2018).

GPS *(Global Positioning System)* is a satellite-based navigation and positioning system.

GSR *(Galvanic Skin Response)* describes the conductivity and electrical properties of the skin and it is related to the arousal level of an individual. A synonym is *EDA*.

HCI *(Human-Computer Interaction)* is a computer science and information technology discipline focused on the study of the interface between technology and users.

HealthKit is a platform/framework in the Apple ecosystem, which allows to access, store and share health information between *apps*. The user can access the own health data through the *Health app*.

HRV *(Heart Rate Variability)* describes the variation in the interval of consecutive heart beats. A low variability is an indication for stress.

IBI *(Interbeat Interval)* references to the time interval between consecutive heart beats. Often the R-peaks of the *Electrocardiogram (ECG)* signal are used and then it is referred to as *RR-Interval*.

IDS *(Initial Design Space)* of Physiological, Wearable Measurement Devices. This design space was developed as a result of two studies presented in Chapter 5 and the evaluations performed in 4. In its initial form it was published at CHI (Hänsel et al., 2018b). It is discussed in detail in Section 6.1.

interpersonal synchrony (also interpersonal coordination, interactional synchronisation) refers to the degree of coordination and 'dynamic and reciprocal adaptation of the temporal structure of behaviours between interactive partners' Delaherche et al. (2012)[p. 351].

IOS Scale *(Inclusion of Others in Self Scale)* is a pictorial scale to measure the closeness created by Aron et al. (1992).

IoT *(Internet of Things)* describes the concept that physical devices — mostly 'smart' and sensing-equipped devices — are connected over the internet to exchange information.

IPAQ *(International Physical Activity Questionnaire)* is an international measure for physical activity assessment by Craig et al. (2003).

IV *(independent variable)* describes the variable or conditions that are controlled and manipulated within an experiment to test for changes in the *dependent variable (DV)*.

Likert Scale is a scale named after Likert (1932), which often used in research and questionnaires, to ask people on their attitude, opinion, and perception towards a topic by letting them rate the topic on an ordinal scale; e.g. a 5-Likert Scale with values from strongly agree (1) to strongly disagree (5).

Mann–Whitney U test is a statistical test method proposed by Wilcoxon (1945) (therefore, it is sometimes called Wilcoxon-Rank-Sum test) and Mann and Whitney (1947). It is usually applied to compare the locations of two independent samples. As a non-parametric test, it is the non-distribution, rank-based alternative to an independent t-test.

MAT *(Mental Arithmetic Tasks)* are mathematical exercises which are shown to induce cognitive mental load and stress.

Microsoft Band 2 is a wrist-worn fitness tracker from Microsoft (2016). It is one of the devices evaluated in Section 5.2 with more details on the device in Section 4.2.

Nexus 10 MK2 is a portable/Wearable biofeedback sensing device from Mind Media. It acted as the reference in Section 5.2 with more details on the device in Section 4.2.

OS *(Operating System)* is the system running on a device which ensures all the basic functionality.

PAD *(Pleasure-Arousal-Dominance)* model from Mehrabian and Russell (1974) is a classification model for emotions.

PANAS *(Positive Negative Affect Schedule)* by Watson et al. (1988) is a scale to assess emotions classified after the model by Watson et al. (1999).

physiological signal include electric signals and biological properties of the human body. This can include *ECG*/heart rate, *Electromyogram (EMG)*, or skin conductance.

PI *(Personal Informatics)* is the field of systems to support the collection and review of personally relevant data (**?**).

Polar H7 is a chest strap to detect heart beats and heart rate by Polar. It is one of the devices evaluated in Section 5.2 with more details on the device in Section 4.2.

PPG *(Photoplethysmography)* describes a non-invasive method to measure the blood flow under the skin by using a light source and sensor.

QS *(Quantified Self)* is the trend of recording parameters of one's life (often with the support of technology and Personal Informatics (PI) systems) to gain knowledge about oneself and reflect.

ResearchKit is an open source framework from Apple, which eases the process of developing research *apps* by providing mechanisms for creating assessments, running experiments with iPhone users or access *HealthKit* data from participants. This allows to create study with a large number of participants.

RR-Interval refers to the time interval between the R-spikes of the *ECG* signal. It marks the time between two consecutive heart beats. This is also often refered to as *Interbeat Interval (IBI)* is the foundation to calculate *Heart Rate Variability (HRV)*.

SAM *(Self-Assessment Manikin)* by Bradley and Lang (1994) is a visual self-assessment scale to assess emotions based on the *Pleasure-Arousal-Dominance (PAD)* model by Mehrabian and Russell (1974).

SDK *(Software Development Kit)* describes libraries and tools provided for software developers to ease the development of applications for certain platforms (e.g., operating systems, hardware systems).

Shapiro–Wilk test is a statistical test of normality proposed by Shapiro and Wilk (1965).

SNS *(Sympathetic Nervous System)* is one of the *Autonomous Nervous System (ANS)*. It is responsible for activating *fight or flight* response triggered by stimuli and stressors.

TestFlight is a distribution tool and platform for beta testing. It allows the app distribution to either internal or external testers aside from the normal Apple *App Store*.

UI *(user interface)* is the space where the interaction between user and device/application occur.

wearable technology and wearable devices (also often called wearables) are, in course of this report, defined as connected technologies which can be continuously worn by a user. They are often equipped with sensors or actuators and often communicate with other devices, such as mobile phones.

Wilcoxon-Signed-Rank test is a statistical test method proposed by Wilcoxon (1945). It is usually applied to compare the locations of two dependent samples. As a non-parametric test, it is the non-distribution, rank-based alternative to an paired t-test.

Publications

Published

Katrin Hänsel, Kleomenis Katevas, Guido Orgs, Daniel C. Richardson, Akram Alomainy, Hamed Haddadi. The potential of wearable technology for monitoring social interactions based on interpersonal synchrony. In: *Proceedings of the 2017 Workshop on Wearable Systems and Applications* (WearSys'18), June 10, 2018, Munich, Germany.

Katrin Hänsel, Romina Poguntke, Hamed Haddadi, Akram Alomainy, Albrecht Schmidt. What to Put on the User: Sensing Technologies for Studies and Physiology Aware Systems. In: *Proceedings of the 2018 ACM International Conference on Human Factors in Computing* (CHI'18), April 21–26, 2018, Montreal, Canada.

Katrin Hänsel, Hamed Haddadi, Akram Alomainy. Demo: AWSense: A Framework for Collecting Sensing Data from the Apple Watch. In: *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (MobiSys'17), June 19-23, 2017 , Niagra Falls, USA.

Katrin Hänsel. Wearable Sensing Approaches for Stress Recognition in Everyday Life. In: *Proceedings of the 2017 Workshop on MobiSys 2017 Ph.D. Forum* (Ph.D. Forum '17), June 19, 2017, Niagara Falls, USA.

Katrin Hänsel. Wearable and ambient sensing for well-being and emotional awareness in the smart workplace. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (UbiComp'16 Adjunct), September 12–16, 2016, Heidelberg, Germany.

Katrin Hänsel, Akram Alomainy, Hamed Haddadi. Large scale mood and stress self-assessments on a smartwatch. In: *Proceedings of the 2016 ACM International Joint Conference on*

Pervasive and Ubiquitous Computing: Adjunct (UbiComp'16 Adjunct), September 12–16, 2016, Heidelberg, Germany.

Katrin Hänsel, Natalie Wilde, Hamed Haddadi, Akram Alomainy. Challenges with Current Wearable Technology in Monitoring Health Data and Providing Positive Behavioural Support. In: *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare* (MOBIHEALTH 2015), October 14–16, 2015, London, Great Britain.

Under Review and Preprint

Romina Poguntke, **Katrin Hänsel**, Hamed Haddadi, Akram Alomainy, Albrecht Schmidt. Developing the Design Space for Physiological Measurement Tools: A Theoretical Foundation of Decision Criteria. Submitted to: *International Journal of Human-Computer Studies* on 15 November 2018

Kleomenis Katevas, **Katrin Hänsel**, Richard Clegg, Ilias Leontiadis, Hamed Haddadi, Laurissa Tokarchuk. Finding Dory in the Crowd: Detecting Social Interactions using Multi-Modal Mobile Sensing. Available on ArXiv, September 2018.

Preface

This thesis is original product by myself, Katrin Hänsel, and none of the text of the dissertation is taken directly from previously published or collaborative articles. Work done in collaboration is stated below:

Chapter 5: The quantitative laboratory study presented in section 5.2 was done in collaboration with Romina Poguntke (Stuttgart University) (as published in Hänsel et al. (2018b)).

Chapter 6: The *Initial Design Space (IDS)* as presented in Section 6.1 was developed in collaboration with Romina Poguntke (Stuttgart University) (as published in Hänsel et al. (2018b)). The follow-up qualitative study on evaluating, refining and extending the initially presented design space with expert and user interviews was done by Romina in collaboration with myself as currently under review in Pogunktke et al. (2018).

Chapter 7: The study on collecting mobile and wearable sensing data during a social event was done in collaboration with Kleomenis Katevas (Queen Mary University London). Kleomenis' research and analysis focused on f-formation detection based on the mobile phone data (see preprint publication by Katevas et al. (2018)) and my work focused on interpersonal synchrony features within the wearable data set (as published in Hänsel et al. (2018a)).

Within this thesis, the APA citation style has been used and reference links within the text are additionally highlighted with a Small Caps formatting. Terms and abbreviations, which can be found in the Glossary and Abbreviations, and names of products have been highlighted with *italic* text for easier identification. Hyperlinks and cross-refereces (including references to figures, tables, chapters, bibliography and glossary entries) which in the electronic version are interactive, are subtly highlighted with grey text.

Acknowledgments

After four years, this journey comes to an end and I would like to acknowledge some of the wonderful people who I encountered on this rollercoaster ride, who helped me to keep continuing on this path and positively shaped this experience.

First and foremost, I would love to thank my advisors Dr Akram Alomainy and Dr Hamed Haddadi for this opportunity to do research in a field I love. With their never-ending patience, guidance and advice, they supported me along this journey. No matter what time of the day or night, they always offered an open ear and the most valuable council to bring me forward. My gratefulness for this opportunity and the chance to work with them cannot be put in words. I would also like to thank Dr Laurissa Tokarchuk for her support, always-open door and ear, and the offering of group meetings which resulted in inspirational discussions. And I would like to acknowledge the EPSRC funding from the Centre of Intelligent Sensing at Queen Mary University of London.

I would like to thank my examiners, Prof Cecilia Mascolo and Dr Tony Stockman, for their valuable feedback and insightful discussions during the viva.

During this journey, I crossed paths with many amazing people who inspired me, guided me, and cheered me up on this long and sometimes lonely road; with some I lost sight, some became acquaintance and some will be life-long friends. I would especially thank Minos, Daniel, Lida, Katja, Aria, Ahmed, Romina, Mohammad, Yousef and the people I encountered throughout the last four years, who have made life and work here at Queen Mary so much brighter and enjoyable.

My gratitude goes to Dr Daniele Quercia and Dr Luca Aiello for the opportunity to broaden my research horizon in the Social Dynamics group at Bell Labs in Cambridge during a summer internship. It was a pleasure and honour working with them and their peers. It was the most amazing summer and I would like to thank Xi, Matheus, Alex, Milan, Louis, Melanie and

Sanja for sweetening my time in Cambridge and being such shiny people; I'll never forget this summer.

Going way back, I must thank the series of serendipitous circumstances leading me to this point. I would like to thank Dr Roderick Murray-Smith for sparking my interest in researching human factors despite my technical computer science background. Without him and people from the Glasgow IDI group, I would not be on this path in my career today; without Lauren I would have never had gained knowledge about this position on a topic exactly reflecting my passion.

Most importantly, I would like to thank my family and friends for their encouragement and support. Papa, I thank you for sparking my interest in technical things and handing a Rubik's cube to my primary-school self. I know you would be proud of how far I have come. . .

Last but not least I would like to thank you, René, for your endless love, patience (heap loads of it), and support during the last decade. You are completing me in a sense I never knew possible. . .

we are the dust
on the stained glass windows
trying to comprehend
the cathedral.

— Rou Reynolds

Part I.

Motivation and Introduction

Chapter 1

Introduction

This introductory chapter presents the motivation, research questions, and contributions of this thesis.

Wearable, sensor-equipped devices – often referred to as wearables – became increasingly accessible during the last years. What began with smart step counters and activity trackers for fitness-enthusiastic insiders, is now leading to a broad market and a growing number of applications. Especially the fitness and healthcare sectors are popular and expected to grow considerably; wristwear, such as smartwatches and fitness bands are amongst the most popular (see Figure 1.1).

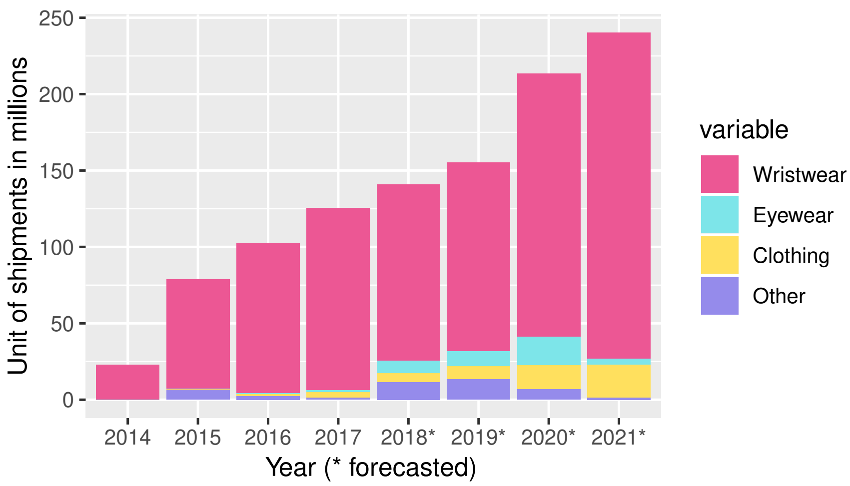


Figure 1.1.: Forecasted units of wearables shipped worldwide from 2014 to 2021 grouped by product category. (Data Source: IDC (2017))

Enabled is this movement by the ongoing trend to shrink computational elements like processors. Moore's Law states that the computational power and the number of transistors on microprocessors double every 18 to 24 months (Moore, 1998). Not just the density of transistors is increasing; new technologies also become less power-consuming due to new energy-saving developments. These developments enable the production of smaller, more energy efficient processors and controllers, which can be integrated into body-worn devices to address the particular requirements of wearable technology and increase the form factor. The development of new energy saving transmission technologies, like for example *Bluetooth Low Energy (BLE)*, enable a more extended and more practical availability of wearable devices while easing the connection and communication with other devices.

The rising social acceptance of body-worn technology is also a driver for the increasing adoption of wearables on the consumer level; it facilitates the application of wearable technology in new sectors of our lives. As mentioned before, the most significant growth potential for wearable technology lies in the health and fitness sector. Ericsson Consumerlab (2015) surveyed consumers on their opinion regarding wearables and discovered that there are high expectations of wearable technology supporting healthier and happier lives. Dunn et al. (2018) forecasted wearable technology to be a big disruptor and to revolutionise health care by offering continuous health relevant data which can be shared not just with health providers and practitioners, but also researchers.

With the rise of these innovative technological gadgets, there comes the enormous potential for researchers to leverage those data sources to gain insight into the individual's health, fitness, behaviour, or mood. It is the first time in history that we can ubiquitously access and continuously monitor signals such as daily steps, sleep, heart rate, and more. These signals and their features can provide insights into people's lives on a new level of granularity and scale.

1.1. Motivation

This work is mainly motivated from a research perspective and focuses on the potential of novel wearable sensing technologies to be leveraged in research. Human-centred research — e.g., in psychology, social sciences, medicine, or *Human-Computer Interaction (HCI)* — since all times faced the problem of observing people in natural, long-term settings opposed to controlled, artificial and lab-based setups. While the latter allows researchers to observe, measure and intervene participants in a structurally and controlled way, laboratory studies are artificial, can lack realism and may lead to hardly generalisable results. However, studying people in the wild and their everyday lives can be challenging. Many confounding factors which have not been considered in the study setup or are impractical to capture can influence and bias the collected data. Novel sensing streams can provide more insights into these factors.

Within the field of *HCI*, 'in the wild studies', as coined by Chamberlain et al. (2012) and Rogers and Marshall (2017), became increasingly valued in the last years. Additionally, Kjeldskov and Skov (2014) argued that it is additionally essential to move away from short-term 'snapshot' studies towards longitudinal studies. Challenges arise with collecting meaningful data on the users within this context; e.g., compliance to fill out daily questionnaires over a longer duration may be problematic.

New, ubiquitous technologies such as mobiles and wearables can support natural, non-intrusive data collection and feedback within the users' everyday life. Wearable devices offer, due to their closeness to the human body, novel sensing means for continuously collecting signals about internal processes. A common example in current wearable devices is heart rate; while it is currently used mainly for gauging fitness related metrics, it has the potential to allow fine-grained insights on mood, stress, or social aspects. Research in the fields of medicine, psychology, social sciences or *HCI* can benefit from a continuous and in-the-wild collection of physiological sensing data from wearables.

Questions regarding the suitability of wearables for this task still remain; there are issues with validity of the data (Alberto et al., 2017; Rosenberger et al., 2016), management of large-scale data in a privacy-preserving and ethical way (Mok et al., 2015; Kelly et al., 2013), ease of integrating devices into the ethical research process (Kreitmair et al., 2017), long-term engagement and retention (Ledger and McCaffrey, 2014), and challenges in terms of analysing large amounts of free-living data (Redmond et al., 2014).

The fast-paced nature of the wearable device market makes the proper assessment of the suitability difficult. Missing standards and loopholes in regulations allow devices to enter the market without their provided data being validated. Further, closed systems, limited access to sensing data, and non-transparent data processing impede easy utilisation of collected sensing data of a device. These factors make the process of choosing the right device for a research setup challenging.

1.2. Research Questions

With the problems and shortcomings highlighted in the last section, the main question arises:

> **How suitable are physiology-sensing wearables for human-subject studies in affective and psychology research?**

This question is not easily and universally to answer due to the multitude of wearable devices, different sensing technologies and the rapid ongoing development. To address this question, an approach from exploring specific devices towards a more generalised conceptualisation of the problem is taken. Finally, guidelines for applying these concepts to assess the suitability of particular and future devices for a research setup are presented.

Exploration. Firstly, it should be explored how sensing data from wearable devices can be obtained. Accessing suitable data for a proposed study is the basis for conducting said research. The question to explore is:

> **RQ 1** *How can research-suitable data be obtained from wearable devices?*

Secondly and after aspects of obtainability of data have been considered, it should be explored how suitable the devices and the obtained data are for affective research to answer the question:

> **RQ 2** *How suitable are current state-of-the-art wearable devices to be applied for measuring stress and affect in the lab and-in-the wild?*

Conceptualisation. The observations made while addressing the first two RQs 1 and 2, shall be generalised and common aspects should be extracted to answer the question:

> **RQ 3** *What are criteria for choosing a suitable wearable sensing devices in a research settings?*

Application. Considering these derived criteria, the following questions focuses on the application of these criteria on novel or future devices:

> **RQ 4** *How can future devices be evaluated for their suitability to be applied in a certain research setting?*

This is done on an exemplary device.

1.3. Contribution and Thesis Outline

This dissertation explores how suitable emerging wearable devices are for being applied to human-subject research. It is first explored, how data can be obtained from wearables and, next, how suitable the devices are based on two studies. Lastly, general design criteria are derived and based on an interview study, these are validated, refined and extended to form a *A Design Space for Wearable, Physiological Sensing in Research*.

While the chapters 2 and 3 build the foundation of this thesis by providing background information and discussion related work, the next chapters focus on answering the previously discussed research questions and contribute in the following ways:

Exploration of Wearable Data Access for Research Applications. To answer RQ 1, Chapter 4 explores how data can be obtained from wearable devices to be used in research. Firstly, requirements for obtaining suitable data for research are discussed. Based on these, four exemplary devices are examined regarding their data access. These four devices, one professional research device and three consumer devices, a) offer physiological sensors related to affect sensing and b) offer very different and representative means to obtain sensing data. It has been discussed how data can be obtained from the three consumer devices, which opposed to the professional-grade *Nexus 10 MK2* do not offer trivial data access. An open-source sensing framework for the *Apple Watch* — called *AWSense* (Hänsel et al., 2017)) — and an app for data collection from the three consumer devices for laboratory studies are presented; both have been used during studies presented in later chapters. Following, the discussion looks at a more generalised view on data access means in modern wearables and discusses approaches from related work. As main contribution, this chapter provides a discussion and summary on issues with wearable data access. It is highlighted what researchers need to consider regarding data granularity and data access when considering a wearable device for their research setup.

Exploration of Applied Affect Sensing in the Lab and Wild with Wearables. To address RQ 2, Chapter 5 explores how suitable wearable devices are for stress/affect sensing in the lab and in the wild. Two studies were conducted. Firstly, the *Apple Watch* was used to collect mood experience samples and sensing data in the wild with the EmoRate app (HÄNSEL ET AL., 2016). Results showed the relationship between sensing data and mood. Further, the four test devices discussed in Chapter 4 were used in a laboratory setup with stress and physical activity conditions. It was shown that the three consumer devices could not show the expected changes in physiological data during stress conditions, but the professional-grade *Nexus* did. However, sensing data of all devices showed correlations with subjectively rated mood. The results have been published by HÄNSEL ET AL. (2018b). As main contribution, this chapter presents what can be learned from these two studies. There are implications in terms of the reliability of the data retrieved from wearables.

A Design Space for Wearable, Physiological Sensing in Research. To address RQ 3, Chapter 6 first summarises the findings from the previous two chapters by deriving an initial Design Space for wearable, physiological sensing in research. This Design Space contains five dimensions representing decision essential criteria for evaluating a wearables appropriateness for a research setup. These dimensions are *Data Reliability*, *Data Richness*, *Data Accessibility*, *Mobility*, and *Comfort of Attachment*. This *Initial Design Space (IDS)* is part of the contribution of work by HÄNSEL ET AL. (2018b). Following, this initial Design Space was evaluated with semi-structured interviews of five research experts. It has been shown that the five initial proposed dimensions can be extended by two additional dimensions (*Trustworthiness* and *Operability*) and several sub-dimensions. This qualitative evaluation is currently under review (POGUNKTKE ET AL., 2018). As main contribution, this chapter summarises the criteria which need to be considered when choosing a wearable sensing device for research.

A Design Space for Wearable, Physiological Sensing in Research. To explore RQ 4, Chapter 7 demonstrates how the Design Space from the previous chapter can be applied to novel and future devices. This is done by demonstrating how the Design Space is applied to the *Empatica E4 wristband* which is used in a research probe on interpersonal synchrony detection.

Part II.

Foundation

Chapter 2

Background

This chapter provides background reading on wearable sensing technologies and their potential to be used in human-subject research. It highlights what wearable technologies are currently available, what sensing means are available within wearables, and how these can be applied for affective computing. A discussion on how wearables can be used by researchers to support health behaviour change was presented at MobiHealth'15 (Hänsel et al., 2015).

In recent years and with the advancement of wearable technologies, more and more sensing and analysis approaches have formed. Modern wearables are equipped with a variety of sensors, such as accelerometers and gyroscopes to detect movements, physiological sensors, ambient light sensors or GPS-enabled positioning. These sensors have been widely used throughout research projects to collect data about user behaviour like physical activity, sports performance, habits, or mental states.

Some of the first commercially available devices for the health and fitness market focused on the sensing of movements with the purpose of tracking physical activity. Simple digital step counters, such as the *Pocket Pikachu* by Nintendo used pedometer sensors to keep track of the steps taken during the day. In a gamified fashion, the user was able to earn 'watts' to feed the virtual Pikachu character on the device; its predecessor — the *Pocket Pikachu 2 GS* — even included an Infra-Red communication port to transfer the collected watts to another player or a GameBoy gaming device. In the following years, and with the development of power-saving communication standards (e.g., *Bluetooth Low Energy (BLE)*), these devices became 'smart' and connected. Early smart wearable devices, such as *Fitbit One* by Fitbit focus solely on step counts, flights-of-stairs and sleep tracking using internal 3-axis accelerometer and altimeter sensors. The collected data is transferred to Fitbit's online database via a proprietary dongle or a Bluetooth enabled smartphone. This technique to offer an online data warehouse and transfer data via the mobile phone acting as a server (Kirby et al., 2016).

With advances in chips and communication technologies, wearables became even connected to the internet themselves, without the need for a mobile phone as a mediator, like, e.g., the *Apple Watch Series 4*. What began with wrist-devices, spread out to other domains and now there are many form factors, with different sensors, applications domains and differing potential to enrich human-subject research.

Broadly speaking any technological device which is worn on the body can be seen as *wearable technology*, including, e.g., regular wrist watches or portable radios (Dunne, 2004). Malmivaara (2009) distinguished two categories of wearable technology: *wearable electronics* and *wearable computers*; while wearable electronics are conceptually linked to the wearer's body to perform a simple set of tasks, wearable computers offer more complex tasks and are re-programmable/reconfigurable. Examples for wearable electronics are, e.g., wrist watches, heart rate chest straps for runners, or wearable glucose monitors and insulin pumps; they are electronic devices which can offer communication capabilities with other devices but are very purposely build for a set of limited tasks. Smartwatches, like, e.g., the Apple Watch or Android Wear devices can be seen as wearable computers which offer a more complex set of tasks which can be extended with custom programs — called apps. What is common for both categories is the closeness to the human body and therefore access to unique sensing data, like movement, physiological or even chemical signals of a person.

The following sections present an overview of wearable device categories, wearable physiological sensing for affect detection, and an overview of related work on the evaluation of wearables in research. Finally, it is discussed what is lacking and where this thesis' work can contributes.

2.1. Wearable Devices and Sensing

Depending on the application field or target population, there are many form factors and device categories for wearable devices. In 1998, Alex P. Pentland forecasted that:

> *'the easiest way to improve intelligence is by augmenting the items we wear all the time — glasses, wristwatches, clothes and shoes — with miniature computers, video displays, cameras and microphones.* (Pentland, 1998)[p. 90]

And his prognosis turned out accurate of the current wearable technology landscape: the most commonly available wearable technologies come in the form of smartwatches or rings, garments, and head-mounted devices.

2.1.1. Wrist Devices

Wrist-worn devices in the form of the modern smartwatches and fitness trackers seem to be the first thing that pops into someone's mind when talking about 'wearables'. Wristwear devices have the largest market share, and it is forecasted to stay this way (Statista). While devices largely differ in their make, formfactor and the integrated sensors, two common categories emerge: *smartbands* and *smartwatches* (de Arriba-Pérez et al., 2016).

Smartbands — the common fitness tracker. With their limited and usually purpose-build smartbands are often focused on providing fitness and health tracking metric; typical examples are Jawbone, early Fitbit devices, or Garmin's Vivosmart line. A non-fitness related example is the *Empatica Embrace*[1] wristband which focuses on stress and seizure detection. Their functionality is often limited to the initial purpose and cannot be extended by apps. They, further, offer limited feedback capabilities and displays. The lack of extensibility of these devices often also limits access to the collected data. One common approach is the transmission of sensing data to online cloud storage via the mobile phone; the data can then often be accessed by the user through an online dashboard or by developers via a provided *Application Programming*

[1] https://www.empatica.com/en-eu/embrace2/ (accessed: 05/02/2019)

Interfaces (APIs) (Kirby et al., 2016). The access of raw data is often not possible and just aggregated or computed metrics in varying granularity are available (de Arriba-Pérez et al., 2016). A second approach is the provision of a *Software Development Kit (SDK)* which allows direct communication with the wearable device to query for sensing data, e.g., the *Microsoft Band 2 SDK* by Microsoft (2016).

Smartwatches — eierlegende Wollmilchsau[2]. Smartwatches, such as the *Apple Watch*, *Microsoft Band* or Android Wear watches differ from smartbands in their extensibility with apps. While the sensors do not differ much — most smartbands and smartwatches contain fitness related movement and heart rate sensing capabilities — the feedback mechanisms and *user interface (UI)* of smartwatches is ofttimes more complex. The extensibility and availability of *SDKs* for wearable development eases the development of third-party apps with custom functionality, *UIs*, and also often access to sensing capabilities. The latter is especially important for researchers and in the context of this thesis.

2.1.2. Hearables

With the miniaturisation of circuits, advances in power-consumption, and latest communication standards (e.g., *Bluetooth Low Energy (BLE)*), hearable such as advanced, smart hearing aids or headphones/earbuds advanced on the consumer market. As an example, the *eSense* open wearable platform by Kawsar et al. (2018) is targeted for researchers allows audio, motion, and *BLE* sensing from the earbud device.

2.1.3. Digital Fashion

There is a prominent *makers' culture* around the topic of tinkering together electronically equipped fashion pieces. Do-It-Yourself (DIY) components in the form of washable and sewable computing boards, sensors and actuators (e.g., LiliPad by Buechley et al. (2008) and ArdaFruit[3] Arduino-based systems) are broadly available and ease the creation of own electronically equipped textiles and garments. Commercially, digital fashion garments comprise smart t-shirts

[2] 'eierlegende Wollmilchsau' is a German expression for a universal tool which tries to address all the problems at once.
[3] `https://www.adafruit.com/category/65` (accessed: 04/02/2019)

(e.g., HexoSkin), trousers (e.g.,Nadi X[4] yoga pants giving feedback), or smart socks (e.g., Owlet[5] socks monitoring the baby's vital signs).

2.1.4. Smart Glasses & Virtual Reality

Google Glass[6] were smart glasses released in 2013; it is the most popular device in the smart glasses category. They offer a small display to present information in an augmenting way without the need to take the eyes off the surroundings. While at the beginning targeted for the consumer market, there were concerns about privacy and the device was in general not well-adapted (Doyle, 2016). Yet, it found application in nice scenarios such as hospital settings (Glauser, 2013) or manufacturing (Swearingen, 2017).

Augmented and virtual reality glasses, such as *Oculus Rift*,[7] *HTC Vive*,,[8] or *Microsoft HoloLense*[9] are head-mounted (and broadly-speaking wearable) devices. They offer either a virtual reality experience without visual access to the surroundings or an augmented and mixed-reality view with virtual overlays on the field of view of the wearer. While these devices can be seen as wearable devices, they will not be considered in this thesis due to their niche applications.

2.1.5. Other body-attachable devices

Depending on the application area or product design decisions, the attachment of wearable sensing devices may happen on different parts of the body. Apart from the device categories above, there have been devices to be attached to shoes (e.g., Nike Plus), within shoes (e.g., Arion smart insoles [10] trousers (e.g., Fitbit Ultra), or fingers (e.g., smart rings).

[4] `https://www.wearablex.com/products/nadi-x-stripes?` (accessed: 02/02/2019)
[5] `https://owletbabycare.co.uk` (accessed: 02/02/2019)
[6] `https://x.company/glass/` (accessed 02/02/2019)
[7] `https://www.oculus.com/` (accessed: 05/02/2019)
[8] `https://www.vive.com/uk/` (accessed: 05/02/2019)
[9] `https://www.microsoft.com/en-us/hololens` (accessed: 05/02/2019)
[10] `https://www.arion.run/product/arion-smart-insoles/` (accessed: 04/02/2019)

2.2. Sensing Signals from Wearables

Wearable devices allow sensing of unique physiological signals due to their closeness to the human body; this distinguishes them from conventional ubiquitous sensing means, such as ambient sensors or mobile sensing. The most common sensors integrated into wrist-wearables are accelerometer sensors in around 86% of devices followed by heart rate and *Global Positioning System (GPS)* sensors which are included in around a third of all wrist-worn wearables (de Arriba-Pérez et al., 2016). In the following an overview of common sensing means included in common wearable devices is presented.

2.2.1. Sensing of Movements

The majority of consumer fitness wearables rely on some form of motion sensing. A typical, power-saving approach is the use of accelerometer sensors which sense the acceleration in one or multiple axes. The retrieved signals can be used to make assumptions on the form, speed, and force behind the movement. Additional sensors, such as gyroscope and magnetometer can add further layers of information about movements and their direction.

Wearable movement data can be used with activity recognition systems to distinguish between activities of daily living (ADL), such as standing, walking, running, climbing stairs (Yang et al., 2010); this is a functionality commonly found in wrist-worn fitness wearables today. More fine-grained activity recognition approaches focus on particular activity domains, e.g., gym exercises. The Atlas wristband 2[11] is advertised as a wrist wearable to detect 100 gym exercises and allows to 'tech' the device new activities. The intensity of physical activity and the calories burned — namely the energy expenditure — are common things estimated by wearables using movement sensors alone or a combination of movement and heart rate signals; but related work has shown that the accuracy of the derived energy expenditure is questionable (Murakami et al., 2016; Rosenberger et al., 2016; Alberto et al., 2017). Another common metric derived from wearable movement data is sleep activity and sleep quality (Shelgikar et al., 2016). Again,

[11] `https://atlaswearables.com/products/atlas-wristband-2` (accessed: 05/02/2019)

the validity of these metrics from consumer fitness wearables is questionable (Rosenberger et al., 2016).

More specialised applications of wearable movement sensors looked into fall detection[12] (Shibuya et al., 2015; Pierleoni et al., 2015), analysing gait patterns (Lee et al., 2007; Sakakibara et al., 2017), respiration movements (Lepine et al., 2016; Jiang and Zhu, 2016), or smoking via arm movements (Akyazi et al., 2017). These approaches are mainly still research and not yet implemented in commercially available off-the-shelf devices. Apart from using accelerometer and gyroscope sensors, stretch sensors which can be embedded in textiles and fabrics have been used to monitor movements. In the future, such sensors could be used for monitoring joint (Huang et al., 2017), body (Yamamoto et al., 2017), respiration (Guay et al., 2017) movements.

2.2.2. Sensing of Physiological Signals

Wearables are equipped with different sensors to pick up physiological signals. Physiological signals can provide insights into the human body and its processes. Apart from medical or fitness applications, e.g., to make assumptions on cardiovascular health or calories burned throughout the day, these sensors can provide information on stress or affect (see next Section 2.3). While in this thesis, especially heart activity, *Electrodermal Activity (EDA)*, and skin temperature are considered, other physiological signals are discussed, too.

2.2.2.1. Heart activity

With the rising interests of consumers in their health and fitness metrics, manufacturers began including sensing means for physiological parameters. A ubiquitous signal is heart rate. Optical heart rate sensors are included in a majority of novel wrist-worn devices. There are two common mechanisms to monitor heartbeats: *Electrocardiogram (ECG)* and *Photoplethysmography (PPG)*.

Photoplethysmography. Smartbands and smartwatches often include optical *PPG* sensors which pick up the fluctuations in blood flow volume in the blood vessels under the skin. This method indirectly picks up heartbeats by sensing the fluctuations in blood volume which are caused

[12] A feature which recently has been included in the Series 4 models of the *Apple Watch* — https://support.apple.com/en-gb/HT208944 (accessed: 04/02/2019)

by the cardiac activity of the heart. *PPG* sensors most commonly use green light emitters and optical sensors picking up the reflected light signal. Wavelengths of green light are particularly suitable for the measurement of blood flow under the skin (Tamura et al., 2014). The peaks in the blood flow are used to detect heartbeats which can be counted to calculate the pulse rate. Reflective *PPG* sensors are very prone to artefacts in the signal due to movement or pressure disturbances whereby pressure can deform the arterial geometry by compression (Tamura et al., 2014). Common measurement sites for *PPG* sensing are the wrist, fingertips (e.g., pulse oximeters), earlobes, or inner ear (e.g., via earbuds). Exemplary wearable devices equipped with optical *PPG* sensing are, e.g., recent Fitbit models (versa or charge 3), *Apple Watch*, *Microsoft Band 2*, or Jabra Sport Pulse[13] wireless earphones.

Electrocardiogram. Another heart rate sensing technology is *ECG*. *ECG* is used to directly sense the electrical activity of the heart. Electrodes placed at specific points on the body and both sides of the heart pick up electrical changes by the heart muscle during the contractions of the heart. Distinct spikes within this electrical signal can be used to detect heartbeats which are used to calculate the pulse. Abnormalities within the *ECG* signal and its features can be used to establish the healthiness or malfunctioning of the heart (de Chazal et al., 2004). The most accurate and gold standard way of measuring *ECG* signals is via the 12-lead setup whereby ten electrodes (4 limb electrodes and 6 electrodes on the chest) are placed at specific points of the body (British Cardiovascular Society, 2010). This setup is mainly used in hospital settings while patients lay down. It is less suitable for non-stationary setups due to the many cables involved. Setups with a reduced number of electrodes, like, e.g., with a battery-powered Holter, can be used in ambulatory and long-term *ECG* recordings. Still, the involved cables may restrict the wearer of the device.

Commercial and consumer wearables for sensing *ECG* signals come in various shapes. There are chest straps (e.g., *Polar H7*), smart garment t-shirts (e.g., Hexoskin), adhesive patches placed on the chest (e.g., Lief[14]), or sensing electrodes included in other wearable devices (e.g., *Apple Watch* Series 4 offers short term *ECG* recordings[15]).

[13] `https://www.jabra.co.uk/supportpages/jabra-sport-pulse-wireless` (accessed: 05/02/2019)
[14] `https://getlief.com/` (accessed: 05/02/2019)
[15] The *ECG* recording on the Apple Watch requires the user to touch the digital crown of the watch with a finger from the other hand to close the circuit

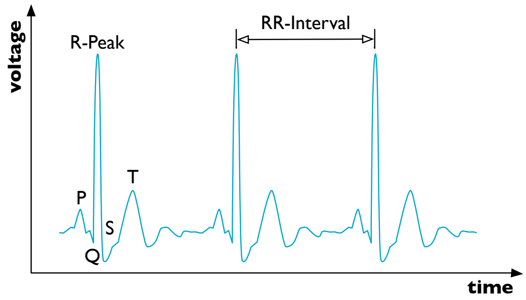
Heart Rate Variability. Both *PPG* and *ECG* allow the calculation of a measure called *Heart Rate Variability (HRV)*. *HRV* is a common *physiological signal* to detect different internal states of the body and it is often used to detect *arousal* and stress states (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996; Thought Technology Ltd., 2010). A healthy heart of a relaxed person does not beat regularly but is influenced by respiration and the *Autonomous Nervous System*; this leads to a constant change of the time between consecutive heart beats. When the person becomes stressed and hormones, such as Cortisol, are set free, the interval between heartbeats become more and more steady, and the variability decreases. Therefore, the *Heart Rate Variability* measures the variability of the inter-beat interval of the heart over time; is the variability of heart rate intervals high, then the person is relaxed. When the person becomes stressed the variability decreases. To calculate the *HRV*, very accurate measurements of the inter-beat interval is necessary. The inter-beat intervals can be derived from both the *ECG* and *PPG* signal; but *PPG* sensors are more prone to noise and good algorithms for filtering are necessary (Ahn, 2013).

Figure 2.1 shows two schematics of a *ECG* and *PPG* signal. The timespan between two consecutive heartbeats is derived based on the peaks in the signals. In case of the *ECG* signal, these peaks are called *R-peaks* and the inter-beat interval is called *RR interval*.[16] The signal (see Figure 2.1b) looks smoother than the *ECG* signal and the heart rate is determined by counting the Systolic peaks; and similarly to the R-peaks of the *ECG* signal, the systolic peaks can be used to calculate the inter-beat time (PP-interval) for the *HRV*.

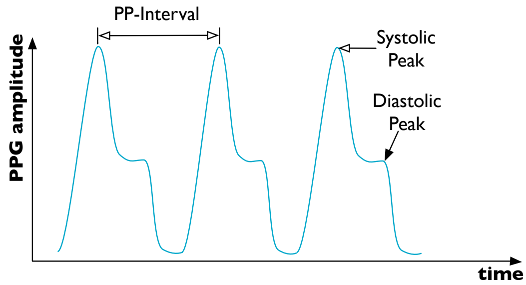
Wearables offering *HRV* metrics are the Garmin VivoFit,[17] Apple Watch,[18] or Hexoskin smart t-shirts.[19]

Heart Rate estimation from Movement Sensors. Apart from using the established *PPG* or *ECG* sensing means, Hernandez et al. (2014) used head-mounted accelerometer data from Google Glass to estimate respiration and heart rate in a controlled laboratory setting; it is unclear how the results would generalise in everyday life scenarios.

[16]The heartbeat intervals are sometimes also referred to as NN-intervals.
[17]`https://support.garmin.com/en-GB/?faq=04pnPSBTYSAYL9FylZoUl5` (accessed: 19/02/2019)
[18]`https://support.apple.com/en-gb/HT204666` (accessed: Accessed: 19/02/2019)
[19]`https://www.hexoskin.com/` (accessed: 19/02/2019)



(a) *ECG* signal with the named waves (P, Q, R, S, T). Waves are used by professionals to discuss certain stages of the heart beat, e.g., the QRS complex marks the activity of the lower heart chamber.



(b) *PPG* signal. The PP-Interval marks the time between consecutive systolic peaks in the blood volume.

Figure 2.1.: Visualisation of the raw signals and signal components from a *ECG* and *PPG* sensor. Highlighted are the signal features that are used to a) calculate the heart rate (R- and P-peaks) and b) calculate the heart beat intervals (RR- and PP-interval) for deriving the *HRV*.

2.2.2.2. Electrodermal Activity (EDA)

EDA describes the changes in the electrical properties of the skin. The most commonly used feature is the skin conductance - also referred to as *Galvanic Skin Response (GSR)*.[20] Changes in the *Autonomous Nervous System (ANS)* can trigger the sweat glands which cause changes in the electrical properties of the skin, e.g., a more moist skin has an increased conductivity. Hence, *GSR* is an indicator of the psychological *arousal* state of the user, and it is linked to emotional and cognitive processes. It has further been shown to be a predictor of epileptic seizures (Poh et al., 2012). The easiness to measure the skin conductance makes it a good tool for *wearable technology*. Poh et al. (2010) presented a small, wearable chip integrated in a wristband to monitor *GSR*. Garbarino et al. (2014) developed a wristband, named E3, to detect *GSR* for real-time biofeedback and data acquisition.

[20] The two terms *EDA* and *GSR* are often used interchangeably, although *EDA* is the umbrella term and describes a broader field.

Common wearables on the marked for skin conductance sensing include: *Microsoft Band*, *Empatica E4 wristband*, and Empatica Embrace.[21]

2.2.2.3. Respiration

Respiration and the patterns of breathing movements are commonly monitored in two ways: by observing the movement of the chest via accelerometers or by sensing the contraction and expansion of the upper body via stretch sensors.

One commercial consumer device for monitoring respiration with accelerometer sensors is *Spire Health Tag*. Holt et al. (2018) evaluated this device and showed an absolute mean error of 1.8 breaths per minute.

2.2.2.4. Brain Activity

An *Electroencephalogram (EEG)* allows the non-invasive sensing of electrical activity within the brain (Davidson et al., 2000). An *ECG* is performed by placing electrodes at certain spots on the scalp to sense the neural activity and brain processes; depending on the placement of the electrodes, different processes can be observed. Often these electrodes — usually wet electrodes — are integrated into *ECG* caps to ease the proper positioning and keep them in place. Common applications for *ECG* sensing are to observe cognitive and affective processes (Davidson et al., 2000), motor imagery *Brain-Computer Interfaces (BCIs)* for paralysed individuals (Quek et al., 2011), or seizure monitoring (Lee et al., 2000).

Consumer devices targeting *ECG* sensing include the *Emotive Insight*[22] 5 Channel *BCI* and *Muse*[23] headband for meditation biofeedback. While these devices are less intrusive than the commonly used *ECG* caps, the social acceptability of wearing these devices in public is still questionable, hence, they are less suited for long-term sensing in everyday life scenarios.

2.2.2.5. Muscle Activity

Similar to *ECG* which senses the muscle activity of the heart, electrodes can be used to detect muscle activity at various places of the human body. This concept is called *Electromyogram (EMG)*. Its applications range from exercise physiology, rehabilitation to biofeedback (Merletti et al., 2004).

[21] Empatica Embrace is targeted for stress and seizure monitoring: `https://www.empatica.com/en-eu/embrace2/` (accessed: 05/02/2019)

[22] `https://www.emotiv.com/product/emotiv-insight-5-channel-mobile-eeg/` (accessed: 05/02/2019)

[23] `https://choosemuse.com` (accessed: 05/02/2019)

A popular wearable for *EMG* sensing is the *MYO* armband; it is discontinued by now (Lake, 2018). It is placed at the lower arm, and it is mainly used as a *user interface (UI)* device for gesture input.

2.2.3. Sensing of Biochemical Signals

The closeness of the wearables to the skin does not just facilitate the recording of *EDA* but also the sensing of chemical compounds in sweat and other body fluids. With the aim to be non-invasive, these technologies are designed as skin patches without the need to penetrate the skin. These chemical compounds can be used to detect and monitor medical conditions and markers such as cystic fibrosis, hypokalaemia, dehydration, blood glucose, muscle lactate, or stress hormones as cortisol (Gao et al., 2016; Bandodkar et al., 2016). Biochemical wearable sensors are still in the research phase, and there are not many known commercially available devices, yet. One example is the *BACtrack Skyn*[24] alcohol-tracking wrist band; it monitors blood alcohol levels using a transdermal sensor measuring alcohol evaporation from the skin.

2.2.4. Sensing of Surroundings

Some wearables not just focus on the recording of human sensing data but also sensing data of surroundings. One example is the *Microsoft Band* which includes a UV sensor (Microsoft, 2016). Other non-yet-released products explore the space of environmental tracking with wrist-wearables (*Sensaris*[25]) or a wearable pendant for allergen detection in food and air (*allergy amulet*[26]).

Other wearable devices, such as *Google glass* or wearable cameras, e.g. Microsoft's SenseCam Hodges et al. (2006), can be used to track visually and analyse the surroundings. This can come handy to provide contextualised information (Clarkson et al., 2000), enable reflection and support memory (Mair et al., 2018), or analyse social interactions (Mandal et al., 2015). Instead of visual sensing, social proximity and encounters have also been tracked using radio-frequency

[24] https://www.bactrack.com/pages/bactrack-skyn-wearable-alcohol-monitor (accessed: 05/02/2019)
[25] https://sensaris.com/ (accessed: 04/02/2019)
[26] http://www.allergyamulet.com/ (accessed: 04/02/2019)

or light sensing by Cattuto et al. (2010) and Montanari et al. (2018) using own custom sensors instead of off-the-shelf devices.

2.3. Affect and Stress Sensing with Wearables

Within this thesis, two research probes on sensing stress and affect with wearables are presented in Chapter 5; following the background work on this topic is presented. The following sections firstly present an overview of psychological measures and scales to model and assess affect before the main part of stress/affect sensing with wearables is discussed.

2.3.1. Psychological Measures of Affect

Apart from the physiological sensing, there are self-rating measures to assess stress and affect. These are mainly based on psychological models and theories. Traditionally, subjective assessment scales were used alone to research affect (Cohen and Pressmann, 2006). Later, these subjective scales became ground truth measures to annotate other signals, e.g., physiological sensing data. Subjective assessment tools have the advantage to be easy to operate for participants and experimenters. They can be administered, e.g., as questionnaires, without long preparation times. Their disadvantage is that they can include bias and that they merely capture snapshots compared to continuous sensing (Kramer, 1991).

There have been various attempts to classify and categorise affect and stress. This section presents an overview of two models and common scales for assessing affect. These serve as reference for the material and approaches to subjectively enquire affect and stress used in the later studies (e.g., Chapter 5).

2.3.1.1. Circumplex Model of Affect

Russell (1980) suggested in his *Circumplex Model of Affect* that affective states and emotions are distributed over two dimensions: activation and pleasantness (also often called arousal and valence). The model assumes that both dimensions are independently present and that distinct emotional categories are characterised by a combination of the two dimensions, e.g., 'excitement' is characterised as both high arousal and valence. An adapted visualisation of this model can be seen in Figure 2.2. This model of affect has been applied within the *EmoRate* app deployed in the study of Section 5.1.

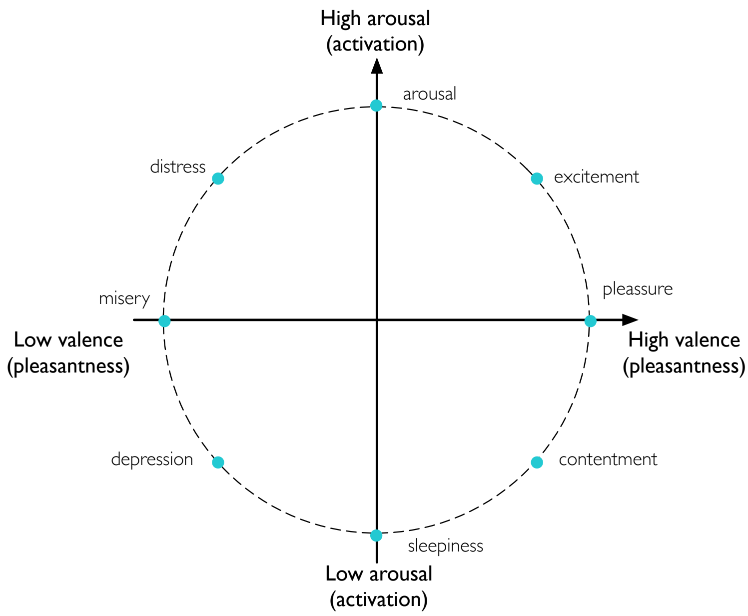


Figure 2.2.: The Circumplex Model of Affect: The eight affect concepts in a circular order (Adapted from Russell (1980))

2.3.1.2. Three-dimensional Model Affect and *Self-Assessment Manikin*

The *Pleasure-Arousal-Dominance (PAD)* by Mehrabian and Russell (1974) is an extension of the two-dimensional model by Russell (1980). It adds a third dimension: dominance. Dominance describes how submissive or dominant an experienced emotion is. A scale for assessing emotions based on the *PAD* model, is the Self Assessment Manikin by Bradley and Lang (1994). This tool quickly and reliably collects the participants' perception of their moods on three dimensions employing pictorial figures to inquires arousal, valence and dominance. The *Self-Assessment Manikin* has been validated with emotional picture datasets and is applied in various studies. Its values have been shown to match emotional and stress responses (Gianaros et al., 2012; Quesada et al., 2014) and the responses were found to be cross-cultural observable (Morris, 1995). A depiction of the *SAM* has been used during the laboratory study in Section 5.2 (see participant material in Appendix E).

Heart Activity. One of the most ubiquitous and accessible measure is the heart activity.[27] During stress responses, the heart starts beating faster to ensure more blood and oxygen being available throughout the body. *Electrocardiogram (ECG)* — the sensing of the electrical activity of the heart — has been widely used as a stress measure amongst various disciplines such as medicine (Hamilton-Craig et al., 2014), psychology (Fisher and Newman, 2013; Schneiderman et al., 2005), and *Human-Computer Interaction (HCI)* (MacLean et al., 2013; McDuff et al., 2016) due to its sufficient reliability. Opposed to directly measure the hearts activity, *Photoplethysmography (PPG)* sensors measure the changes in blood flow under the skin to detect heart beats; it is commonly measured at the ear, fingertips or wrists — as it is commonly done in wrist-worn wearables. It can also be used for stress detection (Sandulescu et al., 2015).

Skin Conductivity. *Electrodermal Activity (EDA)*, which is mostly referred to as the activation of sweat glands and the resulted measurable electrical changes; it is also called *Galvanic Skin Response (GSR)* or skin conductivity (Dawson et al., 2007). Skin conductance can be found in prior work as an indicator for cognitive load (Setz et al., 2010) and stress (Hernandez et al., 2011) also as a "predictor of emotional responses to stressful life events" (Najström and Jansson, 2007).

Skin Temperature. Changes in sweat gland activation (through cooling by evaporation) and changes in blood distribution through the body can indirectly cause changes in the surface temperature of the skin. Skin temperature decreases were used as an indicator for stress (Kataoka et al., 1998; Vinkers et al., 2013).

2.3.1.3. Affect Sensing

The above discussed sensors for stress sensing — heart activity, skin conductivity, and skin temperature — can also be used for a more general sensing of affect and arousal in particular.[28]

While arousal and stress are psychologically distinct from each other, as argued by King et al. (1983), physiological responses are similar. Similar to a physiological stress response,

[27] A detailed overview of heart rate sensing technologies can be found in Section 2.2.2.1

[28] For an overview of dimensional affect models, please refer to 2.3.1

an increased arousal is related to an increased heart rate, increased skin conductivity, and decreased skin temperature (Salimpoor et al., 2009). The pleasantness, i.e., valence, of an affective state has also been shown to be related to increases in the three key measures: heart rate, skin conductivity, and skin temperature (Greco et al., 2017; Witvliet and Vrana, 2007). Torres et al. (2013) argued that *EDA* features are best predictors for arousal, while brain waves in form of *Electroencephalogram (EEG)* features best predicted valence.

2.4. Interpersonal Synchrony Detection with Wearables

Within this thesis, one study to detect interpersonal synchrony with a wearable is presented in Chapter 7; following the background work on this topic is presented.

Interpersonal, social interactions are characterised by complex dynamics, such as taking turns while talking, or mirroring the opposites gestures, mannerisms and facial expressions of the opposite person. During interactions, these nonverbal behaviours or even physiological signals become more similar. This effect is referred to as *interpersonal synchrony* (Nagaoka et al., 2005).

Interpersonal synchrony plays an important role in social development. From an early age on, it is an important mechanism facilitating the formation of a secure attachment between infants and caregiver as well as promoting emotional and cognitive growth for the child (Harrist and Waugh, 2002). Increased synchrony between affective states of mothers and 3-month old infants have been shown to relate to higher self-control in early toddler years; this indicates that early face-to-face reciprocity leads to the development of self-regulatory mechanisms in the child (Feldman et al., 1999). Developmental problems can be reflected in a disrupted and lower magnitude of parental-child synchrony (Lunkenheimer et al., 2018).

According to Vacharkulksemsuk and Fredrickson (2012), in adult life, synchrony mainly functions as tools of rapport. Increased synchrony has also been shown to enhance cooperation between individuals (Valdesolo et al., 2010), promote group bonding (Tunçgenç and Cohen, 2016), or increase positive outcomes in therapy sessions (Marci et al., 2007). Launay et al. proposed that synchrony between is a primal actor in forming group cohesion and promote social bonding through neurochemical mechanisms, i.e. the release of endorphins in synchronised activity.

The detection of *interpersonal synchrony* has been proposed as a basis for various feedback mechanisms. Won et al. (2014) proposed to give real-time feedback on *interpersonal synchrony* to improve, e.g., teaching outcomes (tutors), patient rapport (physicians), or social skills in general. Applications from the field of computer supported cooperative work showed the facilitation of synchrony between players in an *exergames* to promote pro-social behaviour (Park et al., 2013).

2.4.1. Sensing Approaches

While the first works on *interpersonal synchrony* relied on manual annotation and observation whereby behaviourism, movements, or facial expressions were hand-coded (e.g., Bernieri (1988); Cappella (1997)), later works leveraged automated, sensing approaches.

Behavioural Sensing. A large body of work in the field focused on behavioural and coordination and the sensing of movements; this includes sensing of movements of single body parts and the whole body. Common body parts are the head (e.g. Ramseyer and Tschacher (2014)) and hands. Whole-body movements are often measures using visual sensing, e.g. with the Kinect (Won et al., 2014; Fujiwara, 2016), or video cameras (Nagaoka and Komori, 2008). Motion tracking sensors, such as accelerometer sensors, have been used to record movements of various parts of the body. Wristbands are easy and unobtrusively to deploy and have been made use of in studies on interpersonal coordination during dance choreographies by von Zimmermann et al. (2018). Katevas et al. (2015) used mobile phone accelerometer sensors to detect synchronised walking patterns between people.

Physiological Sensing. Apart from the synchronisation of body movements, physiological signals such as heart rate, skin conductivity and skin temperature have also shown to cooperate during interaction (Palumbo et al., 2016). These signals can be easily picked up with modern, wearable sensing devices. During interaction synchronisations in the *Autonomous Nervous System (ANS)* occur. This process is also often referred to as *emotional/stress contagion* (Waters et al., 2014; Elfenbein, 2014). Work by Dimitroff et al. (2017) has shown that the change in these responses can depend on the level of empathy of recipients.

This emotional mirroring and the accompanying synchronisation in responses of the *ANS* can hereby be measured through physiological changes in *Electrodermal Activity (EDA)* (Vanutelli et al., 2017; Karvonen et al., 2016), heart rate (Mitkidis et al., 2015; Konvalinka et al., 2011) and *RR-Intervals* (Suveg et al., 2016).

2.4.2. Synchrony Features and Analysis Approaches

When considering higher synchrony and similarity between behavioural or physiological signals, several features and test approaches can be considered.

Correlation. Common measures of *interpersonal synchrony* contain the correlation of signals. Cross-correlation, whereby the time lag of the signals is considered, is often applied (Delaherche et al., 2012). Related to correlation features, Palumbo et al. (2016) described the following attributes:

Magnitude: The strength of the synchrony often refers to the strength of the correlation between signals.

Sign: The sign describes the positive or negative interlinkage of signals, e.g., positive or negative correlation.

Lag: The time lag can be seen as the time lag applied to one data stream to show reoccurring patterns and correlations.

Direction: The direction of the response describes which signal can be predicted from the other and the direction in which the signals are shifted.

Comparison and Significance Tests. To make assumptions on the significance of synchrony features, they have to be compared to a baseline (Delaherche et al., 2012). A common approach is to use surrogate data. Different methods exist to generate comparable surrogate data, e.g., time-shuffling of data to create a comparison dataset, offsetting one dataset with a significant shift, or associating mismatched partners (pseudo pairs) who did not interact at this point of time (Delaherche et al., 2012).

2.5. Summary

The previous sections showed an overview of the variety of wearable devices currently on the market. These range from popular, well-adapted, broad-purpose consumer devices with high market shares to very specialised niche products. Astonishingly, most of these devices are marketed without proper evaluation and validation. PEAKE ET AL. (2018) claims that just around 5% of devices have been formally validated. Loopholes and missing regulations allow those devices marketed as 'wellness or fitness devices' to enter the market without proper approval (HIREMATH ET AL., 2014). There is a clear benefit for researchers leveraging new data streams about the wearer's health, activity, emotional state and surroundings. While this dhapter presents merely an overview of devices, technologies and potential applications, the next chapter examines related work in detail to highlight how this work contributes to the research community in novel ways.

Chapter 3

Related Work

This related work chapter considers how wearables have been used in research. It is especially shown that there is a lack of a holistic evaluation approach for using wearables — and consumer wearables in particular — in research settings.

While in the last chapter gave an overview of existing wearable devices, their sensing signals and their application in affective computing, this chapter focuses specifically on work related to the main research question of this thesis: *How suitable are physiology-sensing wearables for human-subject studies in affective and psychology research?*

The following sections summarise how wearables have been previously evaluated and what aspects have been missed and are addressed in this thesis.

3.1. Integration of Wearables in Research

With the first commercial devices entering the market and fitness trackers becoming more ubiquitous, researchers began to consider how these novel streams of personal sensing data could be leveraged. This ubiquitous and continuous wearable sensing offers benefits for many human-subject related research areas. Data streams like these were not accessible before; and they could provide new insights for fields such as medicine, psychology and social sciences. But there are challenges to address in terms of ethics, privacy and handling of large amounts of data. The following sections present an overview of work in these fields.

3.1.1. Medicine and Health

The main potential of wearable technology is arguably in the healthcare sector. After all, most devices are marketed to provide fitness, wellbeing and health insights. Some of the main drivers for people to buy wearable devices are health and fitness related (Ericsson Consumerlab, 2015).

Researchers started to consider integrating these novel 'smart' fitness sensors into medical research. Arguably, wearable devices are seen potential revolutionisers of modern health care (Dunn et al., 2018). Hiremath et al. (2014) discussed how wearable *Internet of Things (IoT)* architectures could be integrated in modern health care; they identified key aspects which need to be addressed in future work, e.g., privacy, personalised healthcare, or standardisations. In their opinion, the main benefits lie in decreased cost, increased early detection, and less times that patients need to meet their physicians.

The increasing accuracy and portability of health monitoring sensors supports a less obtrusive data collection and enables long-term health monitoring (Pantelopoulos and Bourbakis, 2010). There are many examples where wearable technology is used successfully in monitoring an individual's recovery from illnesses and rehabilitation (Patel et al., 2012). Applications have focused on, e.g., automated gait tracking and feedback after stroke (How et al., 2013) or knee movement tracking with wearable accelerometers after surgery (Nerino et al., 2013). Providing support for rehabilitation at home can save cost for health providers, promote patient satisfaction and improve quality of life (Wijkstra et al., 1995).

However, the potential of using wearables in healthcare goes way beyond improving care and rehabilitation; continuous tracking of health- and lifestyle-relevant information can support medical research in novel ways by providing access to previously hard to capture longitudinal data (Dunn et al., 2018). The *All of Us*[1] (previously called New Initiative on Precision Medicine (Collins and Varmus, 2015)) is a research program to gather data from millions of people in the US; this data includes data from wearable devices and fitness trackers. The aim is to uncover individual factors in lifestyle, environment and biology towards precision medicine. Other work by Althoff et al. (2017) on large-scale research was using physical activity tracking data from

[1] `https://allofus.nih.gov` (accessed: 05/02/2019)

all over the world to uncover inequality on how the activity is distributed across countries; they identified correlating aspects such as gender, age, obesity, but also factors such as walk-abilities of cities. Medical research studies using wearables, like the *Apple Watch* focus on illnesses like cardiovascular disease (Stanford Medicine, 2018), Epilepsy (John Hopkins Medicine, 2018), or concussions (NYU Langone Health, 2018). These insights on how lifestyle, socioeconomic and environmental factors impact health, can support future applications on prevention and treatments inside and outside the hospital (Dunn et al., 2018).

3.1.2. Psychology, Social Sciences and HCI

Similar to health relevant insights we can gain from longitudinal wearable data, other researchers have focused on finding influences of global events on human behaviour measured through wearable-captured metrics such as steps, sleep or heart rate. Aiello et al. (2018) found disruptions in peoples natural rhythms captured by steps, sleep and heart rate during public holidays and impactful political events; they collected data from over 11,000 Nokia Health wearable users in London and San Francisco over a whole year. Longitudinal, large-scale and geographically distributed studies like these would have been hard to realise without ubiquitous and consumer-based sensing technologies providing data for researchers to tap into.

3.1.3. Human Computer Interaction and Ubiquitous Computing

As mentioned before, wearables are a valuable tool for users to self-track and log aspects of their life within the realm of *Quantified Self (QS)*. Within the *QS* community, data collection mainly happens with the purpose to increase self-understanding and reflect on areas that need improvement in the future (Swan, 2013). Choe et al. (2014) found that health improvements are one of the most stated reasons for self-quantification and especially physical activity is often recorded. They also identified two reasons why self-tracking often fails: the tracking of too many things at once and high tracking efforts. A subdomain of *Human-Computer Interaction (HCI)* research focuses on how self-motivated health tracking and health behaviour change can be supported with improved and supportive design of human-machine interfaces.

3.2. Comparison of Consumer and Professional Devices

A large body of research focused on comparison recent consumer wearables with professional and laboratory devices. These studies focused on comparing various measures of mostly fitness devices to gold standard devices.

The majority of wearable devices in form of fitness trackers or smartwatches include motion sensors; these are commonly used to derive step count, activity types or energy expenditure, i.e., calories burned. Many studies showed discrepancies while testing these devices in the lab (Kooiman et al., 2015) or under various free living conditions (Blythe et al., 2017; Dominick et al., 2016; Ferguson et al., 2015). While validity of step count of most of these devices is high, there are huge differences in validity across devices; however, energy expenditure calculations often lack validity compared to professional calorimetry or accelerometrie devices (Evenson et al., 2015). A study from Shcherbina et al. (2017) showed that non of their tested wearables had an error rate below 20% for energy expenditure estimations.

In terms of affect sensing, physiological signals are of more interest. Many modern wearables are equipped with optical heart rate sensors; but evaluations of these revealed shortcomings when compared to gold standard *Electrocardiogram (ECG)* devices. Wang et al. (2014) found that some devices underestimate heart rate, i.e., Fitbit Charge HR, while others overestimated, i.e., Basis Peak. Dooley et al. (2017) found varying error rates between the three test devices: Apple Watch (1.14% and 6.70%), Fitbit Charge HR (2.38% and 16.99%), and Garmin Forerunner 225 (7.87% and 24.38%); as in other work, error rates were higher under physical activity than rest. On the contrary, Stahl et al. (2016) concluded that the six tested heart rate wearables, e.g., Fitbit Charge HR, Microsoft Band, were accurate with mean errors of 3.3% to 6.2% during running.

Most of this work focused on error rates and error percentages of the consumer devices. Some studies considered consumer devices to relate collected sensing data to subjective affect measures. Zhu et al. (2016) considered motion features from the Pebble smartwatch to infer mood based on the Circumplex Model of Affect.[2] Exler et al. (2016) combined smartphone features with heart rate information from the Moto 360 watch to classify low, medium and high

[2]The Circumplex Model of Affect by Russell (1980) is discussed in Section 2.3.1

arousal with a 68% accuracy. QUIROZ ET AL. (2018) collected heart rate sensing data from the Samsung Gear 2[3] and *Polar H7*. HAO ET AL. (2017) used the *Empatica E4 wristband* for stress monitoring in the office leading to a 81.8% accuracy for binary classification.

Additionally, some commercially available devices already claim to provide stress insights; PEAKE ET AL. (2018) highlighted in their recent survey, that merely 5% of consumer wearables have been formally validated through independent research. They highlight the need for independent evaluation of these devices. Similarly, EVENSON ET AL. (2015) and SHCHERBINA ET AL. (2017) conclude that standardised protocols and documentation of measurement properties can help evaluate future devices.

3.3. Suitability beyond Data Aspects

Beyond the previously addressed data qualities, not much work has been done to evaluate wearables. This work is mainly consumer focused. A common problem in the fitness device market, is users abandoning their devices after a short time; 30% of users stoped wearing their device after six months (LEDGER AND MCCAFFREY, 2014). Researchers have focused on promoting adaption of wearables through increasing device portability or resilience (CANHOTO AND ARP, 2017). KARAHANOĞLU AND ERBUĞ (2011) evaluated the perceived usefulness of wearables for the users; They found that hedonic qualities, e.g., aesthetics or personalisation, are equally important to pragmatic qualities, e.g., robustness or usefulness. KIRBY ET AL. (2016) argues that the main issues regarding the evolution of wearable technology are: electrical power, storage and safety of data, and social acceptability. Similarly, HAGHI ET AL. (2017) crystallised security, power consumption, risk of data loss, and wearability as the main bottlenecks of wearables in healthcare. GRIBEL ET AL. (2016) interviewed academics and IT professionals on the acceptance of wearable technology. From these expert interviews they concluded the substantial role that personality traits have on the consumers' willingness to adopt wearable technology. The

[3] QUIROZ ET AL. (2018) additionally coded a proprietary app that collected motion sensing data from the Samsung Gear 2 smartwatch.

focus again lay on factors influencing the behaviour intentions for consumers to buy and use wearables.

Apart from consumer opinion and acceptance, other researchers evaluated the product design perspective or the design space for applying wearables in special settings, i.e., by healthcare providers. Wang et al. (2015) explored design space of wearables products. They argued that wearable product design has to consider trade-offs from four distinct dimensions: technology, human, form, and interaction factors. Regal et al. (2018) focused on considering the wearable technology in healthcare design space. They interviewed healthcare professionals on their opinion regarding wearable devices in the healthcare settings. Focus lay on wearable devices which could support healthcare providers in their daily tasks, such as communication support or support in critical situations. Hassib et al. (2016) explored needs of users regarding affective feedback from wearables. They considered especially the users' needs in terms of utility, feedback channels, and sharing features.

3.4. Summary and Limitations of Related Work

The main question of this thesis on how suitable wearables are for research. As been shown, wearables are already widely used in research projects from, e.g., medicine or *Human-Computer Interaction (HCI)*. A clear framework on what aspects need to be considered when applying wearable devices in research, is lacking. Further, while research from systems design considered wearable device architectures, e.g., Hiremath et al. (2014) and de Arriba-Pérez et al. (2016), little work is done to examine this field from a researcher perspective. This has been considered in Chapter 4.

A rich body of research focused on evaluating validity of wearable sensing data. This happened in the laboratory of free living conditions; devices were mainly compared to laboratory and gold standard devices. In affective research where physiological signals are conventionally collected with laboratory devices, little work has been done on evaluating the consumer devices suitability.

Beyond this, researcher focused on more consumer-related factors, such as drivers for acceptance of wearables, design of desirable wearable products, or wearable applications ins special domains, e.g., for healthcare providers. There is no clear framework and exploration how wearables, especially consumer wearables, can be utilised for research studies and experiments in human-subject research. Past work mainly considered professional and research perspective to make assumptions on consumer needs and factors, e.g., adoption and acceptance. The researcher's needs are less often considered and discussed. Chapter 6 addresses this and presents a design space on applying wearable devices in research.

Part III.

Exploration

Chapter 4

Accessing Data from Wearable Devices

This chapter discusses how sensing data can be obtained from state-of-the-art wearable devices. The technical details on how physiological data from four exemplary devices, who are consequently used in the following chapters, are presented and discussed. Finally, different approaches on how wearables in general provide data access is summarised, and factors for researchers to consider are discussed.

Consumer wearables are equipped with sophisticated, continuous on-body sensing means; human-subject research can benefit from these to collect sensing data. However, leveraging the data from the devices can be challenging due to the lack of open *Application Programming Interfaces (APIs)*. The majority of these devices do not provide easy or out-of-the-box access to the raw collected data to third-parties. Often they merely offer, if at all, access to snapshots and summarised or processed data (DE ARRIBA-PÉREZ ET AL., 2016). While most *APIs* of current wearables do not grant access to raw sensing data, some smartwatches and wearable sensors offer access.

This chapter explores how sensing data can be obtained from wearable devices to address RQ 1: *How can research-suitable data be obtained from wearable devices?* The approach taken, firstly, considers four specific devices which have been used throughout this thesis. Three of these devices are consumer grade: *Apple Watch*, *Polar H7*, and *Microsoft Band 2*. The *Nexus 10 MK2* wearable on the contrary is a professional devices with medical standard by MIND MEDIA. Secondly, it is explored how wearable devices in general store and process their collected data, and what different access modalities exist. Lastly, a more generalised overview of sensing data

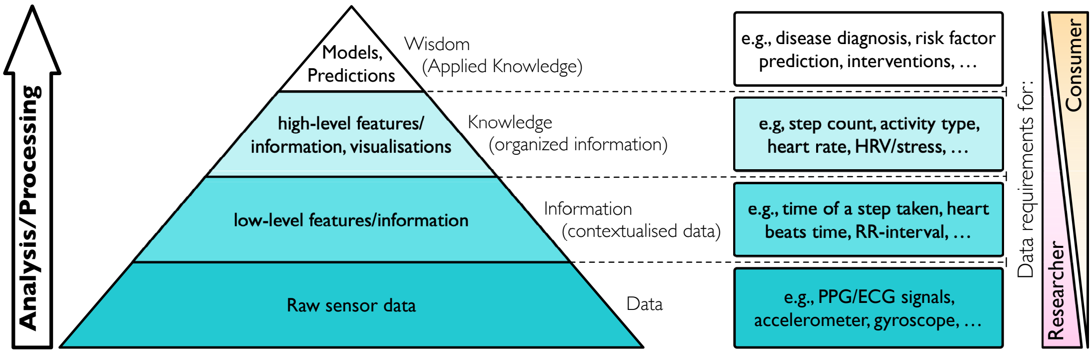


Figure 4.1.: Application of the *Data-Information-Knowledge-Wisdom* (DIKW) (Rowley, 2007) hierarchy to wearable sensing data. With increasing degree of processing, wearable sensing data is contextualised and enriched with meaning. This increasingly contextualised data is usually easier to comprehend for users. On the contrary, more raw and unprocessed data is favourable for professional/research users.

access in wearable devices is presented, and a reflection and discussion on the implications of accessing sensing data for researchers is provided.

4.1. Requirements for Accessing Data from Wearables

When considering consumer devices, these are often designed with the purpose of providing data to the consumer in a summarised and 'digestible' form. To archive this, the sensing data is usually processed and the relevant features are extracted, e.g., the accelerometer data of a device is used to derive step count or even classify the type of activity of a user (Yang et al., 2010). Figure 4.1 shows how wearable sensing data processed in a hierarchical manner and how the context and meaning increases. The *Data–Information–Knowledge–Wisdom (DIKW)* hierarchy (Rowley, 2007) — a theoretical framework on knowledge/information modelling — has been applied to show this.

The *DIKW* hierarchy describes how context and meaning is added to data iteratively by adding analysis and processing. For example, a raw *Electrocardiogram (ECG)* signal[1] does not offer much information; it is merely an electrical signal. Through processing and with the context that distinct spikes (R-spikes) within the signal mark a heart beat, the timings of those heart beats can be extracted. This information can be even more processed to calculate or extrapolate the heart rate, i.e., the number of heart beats per minute, or *Heart Rate Variability (HRV)* which can be interpreted as a stress marker. Additionally, this information can be summarised or visualised to show trends and overviews. Adding sophisticated models or predictions can then be used to make assumptions about future developments, e.g., risk for heart disease, trigger interventions, or to make a diagnosis (Howcroft et al., 2017; Fletcher et al., 2010; Sharmin et al., 2015).

The degree of usefulness of information from a certain level in the hierarchy differs for consumer users and researchers. Most consumer devices and applications deliver higher level information to the wearable device users. Casual users of fitness trackers and smartwatches would not necessarily know how to utilise raw or very low-level data and information from their devices. There is lots of research in itself on how to present wearable device data in a meaningful way to users to support the given goal. For example, researchers explored promoting physical activity from mobile/wearable device data through novel and persuasive visualisations (Consolvo et al., 2008), gamification (Gawley et al., 2016), or just-in-time interventions (Hooke et al., 2016).

Data and information requirements for researchers differ. Usually, data from sensing devices is required in a quantifiable, numeric format for further analysis and processing. But, depending on the research scope, the granularity and required level of this quantifiable information varies. Higher level summaries or visualisations, like they are provided to consumer users, are in most cases not sufficient for researchers. For example, research considering motion sensor data for fall detection requires access to raw accelerometer sensing data. Research focusing on physical activity would rely on higher-level features, such as steps and classified activities – features which are already provided by many wearable device manufacturers. While both research scenarios could be done with raw motion sensing data (and the physical activity

[1]see Section 2.2.2.1 for an exemplary visualisation of the *ECG* signal

features could be calculated by the researcher using existing models, e.g., Chevalier (2016)), the opposite cannot be said for using higher level data for fall-detection. Higher-level data is often provided by device manufacturers without proper documentation of how sensing data has been processed. This raises issues in terms of the appropriateness and validity, e.g., many fitness tracking wearables on the market differed greatly in terms of their reported step count and energy expenditure (Shcherbina et al., 2017; Murakami et al., 2016). Following this, it can be argued that access to lower-level and even raw sensing data would be the most desirable to put the control of the full data analysis and processing in the hands of the researcher.

Within the following section, the data and information access from four exemplary devices is discussed primarily from the professional/researcher perspective.

4.2. Data Access of Four Exemplary Devices

As discussed in the background Section 2.3, physiological sensing data can be applied to detect various psychological and internal aspects of the wearer, such as *Autonomous Nervous System (ANS)* states, stress and affect. The measures most commonly used are heart rate and *EDA*; additionally skin temperature has been used, less commonly. With these aspects in mind, the four devices are discussed in regards to the sensing technologies/data they provide. Two consumer wrist-devices — one of them a programmable smartwatch — have been chosen: *Apple Watch* and *Microsoft Band 2*. These two devices provide optical *Photoplethysmography (PPG)* heart rate sensing. On the contrary, the *Polar H7* is another consumer device which offers *Electrocardiogram (ECG)*-based heart rate sensing. Lastly, one professional, medical-grade sensing device is considered. These four devices have all been used in subsequent research studies (see Chapter 5).

Firstly, the devices are presented and general properties, a short history, and research work where they have been applied is presented. Secondly, the technical details of the data access from these devices is discussed and a short summary is given on what these implies for researchers. An overview of the sensors included in the devices is presented in Table 4.1.

		Consumer			Professional
		Apple Watch	Microsoft Band	Polar H7	Nexux 10
Heart Signals	Heart Rate	●*	●*	●†	●†
	IBI	-	●*	●†	●†
	ECG Signal	-	-	-	●†
	PPG Signal	-	-	-	-
	Skin Temperature	-	●	-	●
	Electrodermal Activity (EDA)	-	●	-	●
	Motion Sensors	●	●	-	-
	programmable *UI*	●	-	-	-

* *Photoplethysmography (PPG)* derived signals; † *Electrocardiogram (ECG)* derived signals

Table 4.1.: Overview of data sources incorporated in the test devices considered. Additional data input from a programmable *user interface* (*UI*) is available for the *Apple Watch*. Two different heart rate technologies (*ECG* and *PPG*) have been indicated.

4.2.1. Device overview and sensors

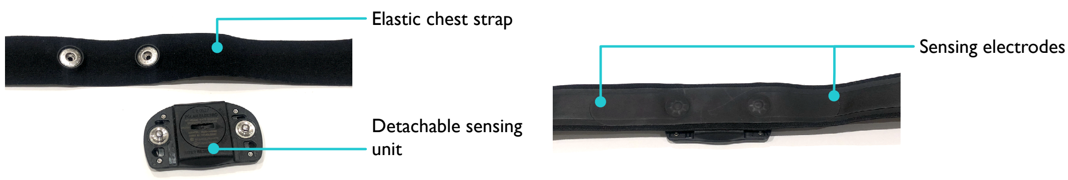
Each of the test devices provides at least one of the relevant physiological sensors (heart rate, skin conductance, skin temperature). Following, an overview of the devices is presented.

4.2.1.1. Apple Watch

The *Apple Watch* is a wrist-worn wearable by Apple which was first released in 2015 (Tech Crunch, 2015). It is a smartwatch which can be used with iOs smartphones. As s a smartwatch, it offers a rich *user interface (UI)* which can be extended with *apps*. Beyond the optical, *PPG* heart rate sensor, it also includes motion sensors in the form of accelerometer, gyroscope and magnetometer. A *Global Positioning System (GPS)* sensor has been added as an option of the second generation devices; it is a standard sensor in all watch models of the third and fourth generation. The fourth generation introduced an *ECG* sensor in the device;

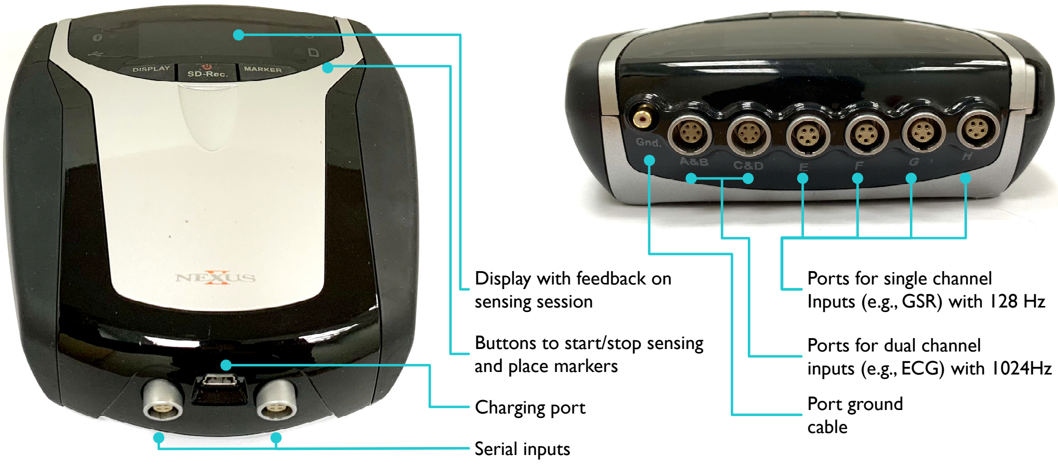
(a) Microsoft Band 2

(b) Apple Watch



(c) Polar H7

Figure 4.2.: Overview of consumer test devices



(a) Nexus main device

(b) ECG electrode cables (2 sets) and ground electrode (white)

(c) Snap-on, pre-gelled electrodes for ECG

(d) Temperature sensor to be attached to the skin

(e) Finger electrodes for GSR sensing

Figure 4.3.: The *Nexus 10 MK2* and its sensors.

this sensor is not yet accessible to developers and is not currently available for users outside the United States or the European Economic Area (Wired, 2018; Apple Inc., 2019a). Figure 4.2b shows the device, its components, and its visible sensors. The *Apple Watch* is considered as a test device in this thesis, due to its popularity and high market share, programmable *UI* (later utilised in a study in Section 5.1) and its shown validity in previous, related work; several studies found the *Apple Watch* superior to other wrist devices in terms of heart rate error and correlation with the reference device (El-Amrawy and Nounou, 2015; Shcherbina et al., 2017; Wallen et al., 2016).

4.2.1.2. Microsoft Band 2

The *Microsoft Band* fitness tracker is another wrist-worn wearable, with cross-platform support for Windows, iOS, and Android phones (Figure 4.2a). It was released in 2015, and it is special with regards to its available sensors. Apart from the relevant physiological sensors (*PPG* heart rate, skin temperature, and *EDA*), it also offers motion sensors, UV sensing, and a barometer (Microsoft, 2016). It is worn around the wrist and offers a simplistic *UI* which is extendable with simple functionalities, such as presenting notifications and dialogs. It is not able to be extended with more complex apps. The *Microsoft Band* is considered here and in the following user studies, due to its rich sensor set. It is one of the few consumer wearables incorporating a skin conductance and skin temperature sensors which are relevant measures for physiological arousal. The device was discontinued in 2016 and all the services were terminated in May 2019 (Warren, 2019).

4.2.1.3. Polar H7

The *Polar H7* is a chest strap wearable with *Bluetooth Low Energy (BLE)* connectivity (Figure 4.2c). Its sole integrated sensor is a heart rate sensor based on *ECG* technology. Due to its *BLE Generic Attributes (GATT)* standard-conform connectivity, it is compatible with any *BLE*-enabled device, e.g., modern smartphones, laptops, and even smartwatches (Caldwell and Filipowicz, 2018). *ECG* sensing is performed through two electrodes integrated in an elastic strap. The detachable sensing unit contains a 3-volt lithium coin cell battery slot for a battery life of up to 200 hours (Polar Support, 2017). To save energy, the device turns it self on when skin contact is established and off once skin contact is lost; there is no feedback on the device to indicate this. The *Polar H7* was chosen as a test device due to its low cost *ECG* technology and cross-platform availability. Heart Rate chest straps have been shown to have a high accuracy (Gillinov et al., 2017) and have been used as criterion devices in related work (Stahl et al., 2016).

4.2.1.4. Nexus 10-MK2

The *Nexus 10 MK2* by Mind Media is a professional, medical-grade sensing device. It is a wireless and wearable[2] device which is targeted for biofeedback applications and psychological research. It offers a range of input channels for various sensors which can be plugged into the device (Figure 4.3), e.g., *ECG* heart rate, skin temperature, *Galvanic Skin Response (GSR)*. A small display and buttons allow the operation of the device, e.g., turning the device on/off or placing timestamp markers. The sensing input channels can be configured using the MindMedia (2017) *BioTrace+* software which also allows real-time data visualisations from the bluetooth connected *Nexus*, start/stop of the recording and marker placement functionalities.

4.2.2. Data Properties and Access

The previously presented devices not just offer differing sensors and sensing technologies (i.e., *PPG* vs. *ECG* heart rate) but they also offer differing granularities and access means. These are discussed below per each device. The focus lies on obtaining data in a research-usable way, i.e., numeric vales. This access is named 'pro' access for researchers in the following sections. Since three devices (*Apple Watch*, *Microsoft Band*, and *Polar H7*) are consumer devices, the normal data access means meant for consumer 'users' are briefly presented, too.

An overview of the devices' sensors, sampling rates and measurement units can be found in Table 4.2.

4.2.2.1. Apple Watch

While the *Apple Watch* offers optical heart rate, movement and location sensing, the means to access the sensor data differs. Computed health features, such as *resting heart rate*, *floors climbed*, or *exercise minutes* are stored in the *Health* app on the iPhone. It is accessible through *HealthKit* or through the Health app in form of visualisations. Data in the *Health* app is backed up and synchronised across the iOS devices through the iCloud online storage (Apple Inc., 2019b), but it is not accessible directly from this cloud storage. Raw

[2]The device is wearable in a sense that it can be attached to clothing using a belt.

		Sampling Rate	**Unit**
Nexus	heart rate	32 Hz	bpm
	ECG	256 Hz	micro volts
	EDA - skin conductance	32 Hz	micro siemens
	skin temperature	32 Hz	°C
Apple Watch	heart rate	not specified	bpm
	raw acceleration (x,y,z)	100 Hz	G†
	attitude (x, y, z),	100 Hz	G
	rotation (yawn, pitch, roll),	100 Hz	radians
	gravity (x, y, z),	100 Hz	G
	linear acceleration (x, y, z),	100 Hz	G
	magnetic field (x,y,z)	100 Hz	micro tesla
Microsoft Band	Heart Rate	1 Hz	bpm
	RR Intervals	value changed	seconds
	EDA - Skin Resistance	0.2/5 Hz	koms
	Skin Temperature	1 Hz	°C
	acceleration (x,y,z)	62 Hz	G
	angular velocity (x,y,z)	62 Hz	$°/s$
Polar	Heart Rate	1Hz	bpm
	RR Intervals	value changed	1/1024 seconds

† gravitational force G (9.81ms^{-2} on the earth)

Table 4.2.: Overview of the sensor sampling rate and units of the devices used for the lab study. The information has been retrieved from the manufacturers' manuals and documentation (Microsoft, 2016; Polar, 2018; MindMedia, 2017; Apple Inc., 2018).

motion sensing data from the*Apple Watch* is just accessible through the *CoreMotion* *Application Programming Interface (API)* on the watch (Apple Inc., 2018). A schematic depicting the access means are depicted in Figure 4.4.

HealthKit Access. *HealthKit* is an *API* provided with the iOS *Software Development Kit (SDK)* to read and write health data to the *Health* database on the iPhone (Apple Inc., 2019b). This database contains health metrics of the iPhone owner, such as vital markers, fitness activities, or nutrition. Any iOS app can read and write to this database after the user's permission is granted. Higher level features derived from the the *Apple Watch*'s sensing data (e.g., step count derived from the motion sensor, or heart rate from *PPG* sensor) are stored in this database. This data can be accessed programmatically (from within an app) through Apple's *HealthKit API* either on the phone or on the watch; provided that user's permission has been granted for the specific app. No special *SDKs* or libraries are needed since the *HealthKit API* is integrated in Apple's standard *SDK*. Additionally to real-time data, *HealthKit* also allows access to historic health data.

In summary, access to the health related data and higher level features (including heart rate) in most cases needs to be performed either programmatically or with actual access to the device. For remote studies and data collection, an app needs to be built which incorporates the data collection mechanisms.

Raw Motion Sensor Access. The raw motion sensors, namely accelerometer, gyroscope, and magnetometer can be accessed via Apple's standard *SDK*. Apple's Core Motion *API* provides functions to access motion sensor data in the form of raw acceleration data and aggregated device motion features (Table 4.2). Access to this data is real-time and no historic data can be accessed. While these functions allow the access of the sensor data on the watch itself, there are no readily available means to transmit this data to the phone.[3]

UI Data. The *Apple Watch* smartwatch is the only device of the four test devices to offer a sophisticated programmable *UI*. Programmable *UIs* can become data sources in themselves, e.g., by collecting questionnaire data or user feedback. The *Apple Watch*'s *UI* has been used to collect experience samples during a study presented in Section 5.1.

[3]The *AWSense* framework presented in Section 4.2.3.1 provides easier data access

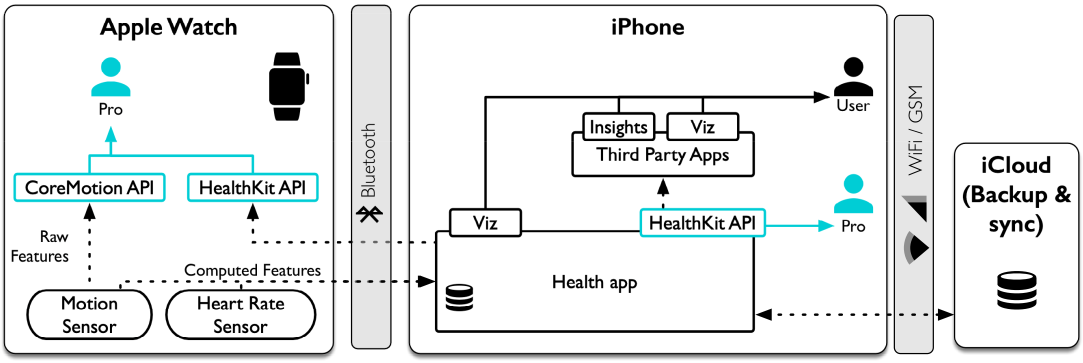


Figure 4.4.: Schematic figure of data access on *Apple Watch*. Data can be accessed programatically by professional users using the provided *APIs* on the watch or phone. computed health features (e.g., steps, burned calories, heart rate) are stored in the *Health* app on the phone and can be accessed via the *HealthKit API*. Raw motion sensing features can just be obtained from the watch itself using a programatic approach.

4.2.2.2. Microsoft Band 2 The sensing data from the *Microsoft Band* can be obtained through different means. Sensing data from the device is generally transferred to the paired phone via *BLE*. There, it is stored in the Microsoft Health companion app on the connected phone. Computed features, e.g., steps, heart rate, burned calories, can be accessed by the user in form of summaries and visualisations. Additionally, this computed data is transferred to the Microsoft Online Health service where the data is stored in the cloud. Users can access summaries and visualisations via a web dashboard. Online *APIs* are provided to allow developers access to this higher-level data; this function could be leveraged by researchers. This approach is not unique to the *Microsoft Band*. It is the common approach to transfer data from modern wearable fitness trackers to a proprietary mobile phone app via Bluetooth. Often, this mobile phone acts as a gateway to transfer data to a cloud service for storage from where it can accessed via a web dashboard or an *API*. In line with *Internet of Things (IoT)* principles, this data can be accessible to third-party services via this offered online *API*.

This approach is followed for fitness wearables from, e.g., Fitbit,[4] Garmin,[5] or Xiaomi.[6]

[4] `https://www.fitbit.com` (accessed: 10/05/2019)
[5] `https://buy.garmin.com/en-GB/GB/wearables/wearables/c10001-c10002-p1.html` (accessed: 10/05/2019)
[6] `https://www.mi.com/global/` (accessed: 10/05/2019)

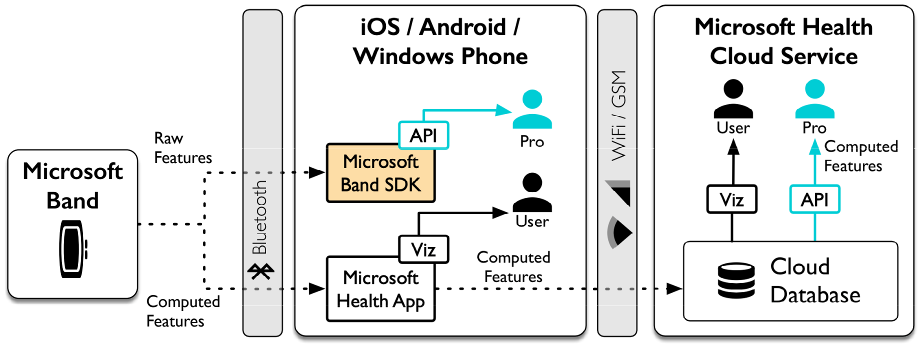


Figure 4.5.: Schematic figure of data access on the *Microsoft Band 2*. User access to higher-level health features in form of visualisations and summaries is provided through the Microsoft Health app on the paired mobile phone or a web dashboard through the Microsoft Health Cloud service. The cloud service also offers an online *API* for accessing/downloading the data. Professional access to raw sensing features can be obtained on the mobile phone via *SDK*.

Additionally, the *Microsoft Band* offers data access via a provided *SDK*.[7] This *SDK* was provided by MICROSOFT (2016) for iOS, Android and Windows Phone, and it can be integrated in mobile phone *apps* for these platforms. This *SDK* allows direct, programatically access to real-time, raw sensing features (Table 4.2) which can be leveraged by pro users, e.g., researchers. It does not allow the access to historic sensing data. The provision of a mobile *SDK* for sensing data access is not common amongst wearable device manufacturers and is a speciality of the *Microsoft Band*.

4.2.2.3. Polar H7

The *Polar H7* is a representative device for *BLE*-enabled heart rate sensing straps. As such, it can be paired and connected to a multitude of *BLE*-enabled devices, e.g., mobile phones, modern computers, smartwatches, etc. It's connectivity is based on the *GATT* Heart Rate Service standard (BLUETOOTH SPECIAL INTEREST GROUP, 2011). While the manufacturer does not offer an *SDK*, a developer documentation with sample source code exists (POLAR, 2018). To access sensing data from the *Polar H7* the paired device's *BLE* programming layer *API* has to be used to allow access to real-time sensing data from the *Polar H7*. This offers pro users the opportunity to connect the *Polar H7* to a range of devices and integrate it in various research

[7]The *SDK* is not advertised and available on the manufacturers website any more, since it is discontinued.

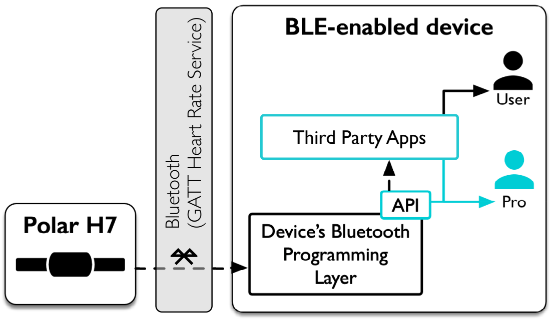


Figure 4.6.: Schematic figure of data access on the *Polar H7*. Access to the sensing data can be obtained through the paired device's *Bluetooth Low Energy (BLE)* programming interface or third-party apps.

applications. The access through the device's bluetooth layer is also used by third-party apps which connect to the *Polar H7*, e.g., *Strava*[8] fitness app or *HRVLogger* by Altini (2013). These are often meant for users of the device, e.g., for sophisticated fitness tracking, but can also be leveraged by researches, providing they allow suitable export options. E.g., *HVRLogger* which allows the download of the raw sensing data and additionally computed *Heart Rate Variability (HRV)* features. A schematic figure of data access from the *Polar H7* is depicted in Figure 4.6.

Other examples of Bluetooth-enabled devices relying on the *GATT* Heart Rate service standard exist. These devices offer the same access means. Other *ECG* heart rate straps are, e.g., the successor Polar H10,[9] or Whaoo TICKR range.[10] Some *PPG*-based, optical heart rate monitors also support this standard, too, e.g., Mio AlPha.[11]

4.2.2.4. Nexus 10

The *Nexus 10 MK2* is a laboratory device with professional application and its main purpose is provide medical sensing data in a professional/research-usable form. It offers two means: internal recording on an SD card and real-time streaming to a Bluetooth-connected Windows PC. In both cases, the device has to be firstly configured through the proprietary *BioTrace+* software by MindMedia (2017). The configuration includes the type of sensors plugged into the *Nexus*, as well as, sampling rates of the sensors. The software also

[8] https://www.strava.com/ (accessed: 08/05/2019)
[9] https://www.polar.com/uk-en/products/accessories//h10_heart_rate_sensor (accessed: 10/05/2019)
[10] https://uk.wahoofitness.com/devices/heart-rate-monitors (accessed: 10/05/2019)
[11] https://www.mioglobal.com/ (accessed: 10/05/2019)

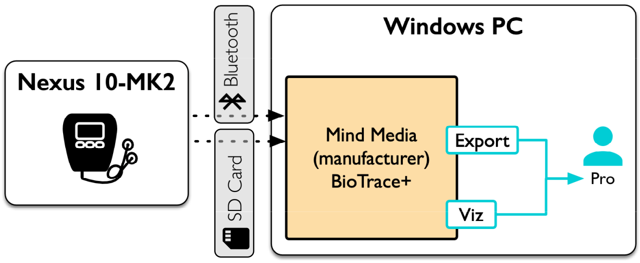


Figure 4.7.: Schematic of data access on the *Nexus*. Data visualisation and export are available through the MindMedia (2017) *BioTrace+* software for either streamed data (Bluetooth) or data stored on the device-internal SD card.

needs to be used in both cases — internal recording or streaming — to access the data; the data on the SD card is stored in a proprietary format and cannot be read out-of-the-box. The *BioTrace+* software offers visualisation of the data either from the real-time streaming or from an SD card recording. In both cases, the data can be exported in various formats, e.g., *Comma Separated Values (CSV)* format. An advantage of this technique is, that data can be seen in real-time to a) allow manual inspection on the face validity of the data, and b) support applications where real-time access is necessary, e.g., biofeedback. The recording on the internal SD card can be advantageous to have a back-up of the data (in case of real-time recording on the PC) or can be used during ambulant settings where a connection to the PC is not possible. A similar approach of real-time streaming to a proprietary application and additional internal recording on an SD card, has also been applied in Shimmer devices[12] and the *Empatica E4 wristband*. Shimmer and Empatica are both companies providing professional sensing tools for research application.

4.2.3. AWSense and LabExperiment App

As shown above, the four devices have very differing approaches to offer data access to professional users (researchers) and consumer users. The access to research usable, numeric data is just supported out of the box by the *Nexus*. The three consumer devices do offer eased access to computed fitness features to their users in form of summaries or visualisations. Access to numeric data, which is processable for researchers, is scarce. For the research probes presented

[12] `http://www.shimmersensing.com/` (accessed: 10/05/2019)

in the next Chapter 5, a library and an *app* for eased data collection were created: *AWSense* and LabExperiment app. These two *apps* are open-source available, and they demonstrate how data from the three consumer test devices can be obtained.

4.2.3.1. AWSense

As discussed before, access of raw motion sensing data from the *Apple Watch* is not possible with conventional means. Accessing raw motion sensor data from the device has to be performed on the wearable itself (see Figure 4.4). This bears the problem of transferring and downloading the data from the device for a researcher to access. Since there is no native solution, the *AWSense* framework[13] has been developed. It was demonstrated at MobiSys'17 (Hänsel et al., 2017). The *AWSense* libraries ease and unify the access to motion and heart rate sensors on the watch and their transfer and storage on the iPhone. More information on the architecture and demo applications can be found in Appendix A.

Battery Implications. The impact of continuous sensing and data transmission on the *Apple Watch*'s battery was evaluated; data recordings in different modes were undertaken while the battery levels of the wearable where logged. This was done using an iPhone 6 with iOS 9.3.1 and an *Apple Watch* Series 1 with watch OS 3.2.2. The trends in battery life are depicted in Figure 4.8.

The benchmark was performed with 4 different modes:

- *heart rate (without transmission)*: to force continuous heart rate monitoring on the *Apple Watch*, a workout was started on the device to record the heart rate.[14] The heart rate was not transmitted using *AWSense* but Apple's native synchronisation to the iOS *Health* app still occurred.

- *all sensors (without transmission)*: Continuous heart rate sensing and collection of accelerometer and device motion samples (50Hz) was performed. The data was not transmitted to the phone.

[13] `https://github.com/MiezelKat/AWSense` (accessed: 13/05/2019)

[14] In workout mode, the *Apple Watch* records heart rate samples approximately every 3 seconds, whereby normal heart rate samples are collected at arbitrary times (Apple, 2016).

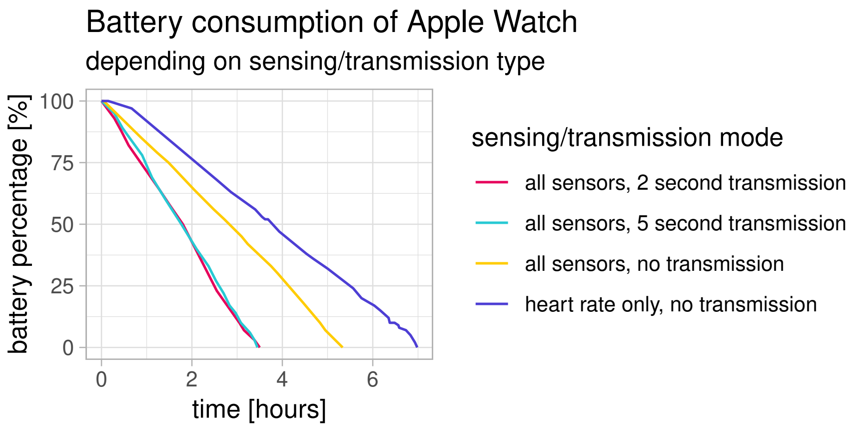


Figure 4.8.: Battery consumption of varying data sensing and transmission modes of the *AWSense* library. The modes were: heart rate without transmission (continuous heart rate was recorded but not transmitted to the phone); all sensors, no transmission (continuous heart rate data, accelerometer and device motion (50Hz each) were recorded but not transmitted to the phone); and all sensors, 2 second/5second transmission (continuous heart rate data, accelerometer and device motion (50Hz each) were recorded and transmitted in the given time interval (2 or 5 seconds)).

- *transmission every 2/5 seconds*: Continuous heart rate sensing and collection of accelerometer and device motion samples (50Hz) was performed. The data was transmitted to the phone every 2 or 5 seconds.

The total drain of the battery in when heart rate was recorded (without transmission) took 6:59 hours while recording heart rate, accelerometer, and device motion without transmission to the phone drained the battery in 5:20 hours — 24% faster. Transmitting the data to the phone instead of just recording it, resulted in a 34-35% faster drain. It took 3:30 hours for a 2-second transmission interval and 3:27 hours for a 5-second transmission interval.

This shows that additional motion sensing data collection opposed to just heart rate data shortens the battery life of the *Apple Watch* significantly. Adding additional data transmission, further drains the battery. For continuous and long-term data collection, these aspects have to be taken into account.

4.2.3.2. LabExperiment app

The LabExperiment app[15] was developed by the thesis author and serves as an example on how data collection from multiple wearables can be unified for an application during laboratory research. It combined data accessibility functionalities for *Apple Watch* (using *AWSense*), *Microsoft Band*, and *Polar H7* using the above discussed data access means. The main purpose of the app was to provide a combined and unified data collection of the three consumer devices, and it has been utilised in the later presented, laboratory experiment (Section 5.2). With this purpose in mind, the LabExperiment app provides the collected sensing data in a research-appropriate format — the *CSV* format which is a human-readable and broadly-supported data format. Beyond that, the app offers to automatically label the data with a custom chosen participant ID and unique session ID to avoid that data is mixed up.

The *UI* shown during data recoding presents researchers with an overview of the current session, such as, the elapsed time or the real-time sensor readings. The later can be used by the researcher to check the face validity of the sensor readings and identify problems early-on, e.g., during the laboratory experiment, the *Polar H7* lost contact with the skin and turned itself off which resulted in no readings shown. Further, researchers can place time markers to e.g., mark the begin/end of an experiment condition. Screenshots and an architectural overview of the app are presented in Appendix Appendix B.

4.2.4. Summary

In this section, four exemplary wearable physiological sensing devices were discussed in terms of access to the sensing data. All devices followed different approaches on providing data access and the presented data accessibility architectures can also be found in other wearable systems. For some architectures, e.g., in the *Apple Watch* and *Microsoft Band*, data access differs for users and professionals.

[15] source code available on Github `https://github.com/MiezelKat/LabExperimentApp`

4.3. Wearable Data Flow Architectures in Consumer Devices

Moving away from the four test devices presented in the last section, approaches have been undertaken to discuss data access in general. Especially, the consumer devices are considered here, since professional devices for sensing data collection, e.g., the *Nexus 10 MK2* or Shimmer sensors, are purposefully built for data access by researchers/professionals. Most modern consumer wearables, e.g., in form of smartwatches and fitness trackers, are not meant to be used on their own. One aspect which makes them 'smart' and distinguishes them from ordinary watches or pedometers, is they connectivity to other devices. They fit well in the *Internet of Things (IoT)* domain of devices connected to other devices and the internet. This not just aides the distribution of data for its availability, e.g., access of health data from a web dashboard, but also supports the processing of the data where more resources are available. To save energy and resources, computational-heavier tasks are offloaded to other hardware and systems, e.g., the cloud (Seneviratne et al., 2017).

To reach the internet and 'the cloud', wearables usually use gateways such as mobile phones (Hiremath et al., 2014; Kirby et al., 2016); the prevalent connection technology is *Bluetooth Low Energy (BLE)* to pair with this gateway. Recently, smartwatches direct connections to the internet through WiFi or *embedded SIM (eSIM)*, e.g., the Apple Watch Series 4 or Samsung Gear S2 models.[16] As discussed in the previous sections of this chapter, but also summarised by de Arriba-Pérez et al. (2016), data access can be provided on each of these three nodes: the wearable device, the gateway device (e.g., mobile phone), or the cloud data warehouse.

Figure 4.9 summarises data access strategies prevalent in modern consumer wearables; the data flow and processing have been extensively discussed and summarised by Hiremath et al. (2014). The collected sensing data is minimally processed and temporarily stored on the wearable. This storage often occurs just until the data can be transferred to the gateway device due to the limited memory and processing resources on the wearable device (de Arriba-Pérez et al., 2016). Data access directly on the wearable device is dependent on the device being programmable and

[16] `https://smartwatchhelp.com/smartwatches-with-esim/` (accessed: 06/06/2019)

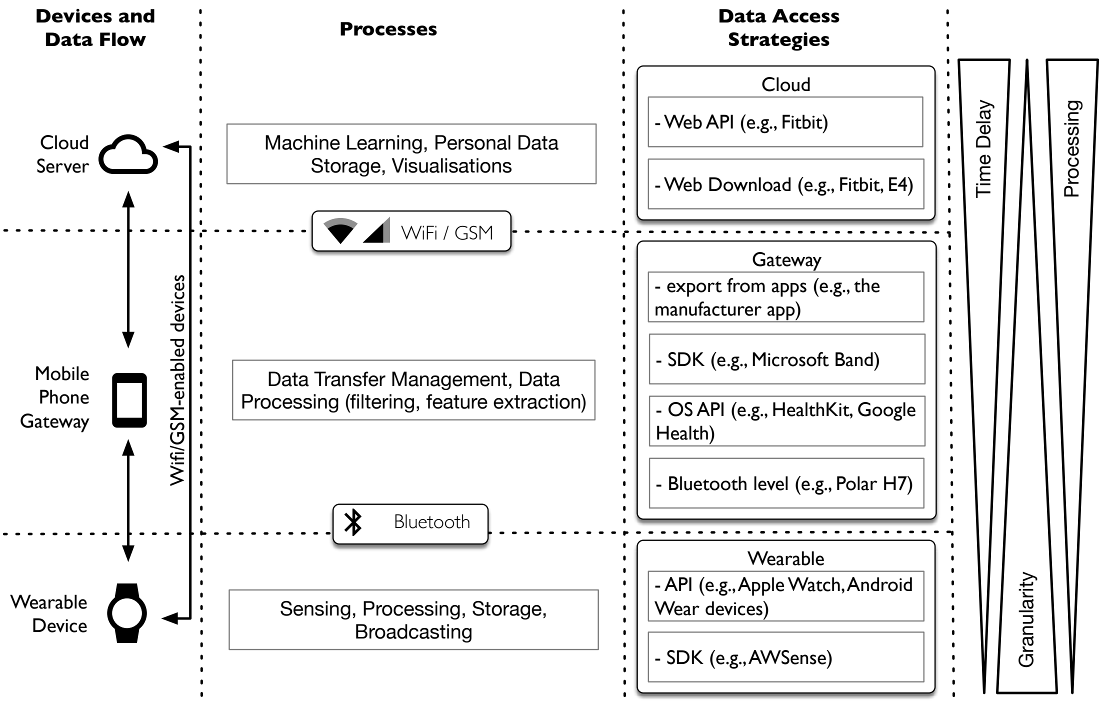


Figure 4.9.: This graphic visualises the data flow, data processing mechanisms, and access strategies for wearable device data with examples. The sections of 'Device and Data Flow' and 'Processes' have been adapted from Hiremath et al. (2014). The 'Data Access Strategies' were concluded from de Arriba-Pérez et al. (2016) and the evaluation of the devices from Section 4.2.

access though the manufacturer's *Application Programming Interface (API)* or *Software Development Kit (SDK)*. Potential applications where researchers would want access directly on the device are, e.g., wearable machine or deep learning applications (Dargazany et al., 2018).

Data access on the intermediate gateway device also depends on the means provided by a) the mobile device *Operating System (OS)*, b) the wearable device vendor, or c) third parties. Depending on wearable device vendor, the collected health features can be available from mobile personal health storages, e.g., Google Health or Apple's Health storage (Dameff et al., 2019). These health storages allow access to this data through *APIs*. Another approach is that device manufacturers support data access on the phone by providing an *SDK* or using open standards, e.g., open *BLE* standards. Lastly, data can be exported from the device vendor's app or third party apps which leverage one of the above means to access the data.

Data transferred to the cloud server usually serves the purpose to be accessible by the user in a form of web dashboard with summaries and visualisations (Kamišalić et al., 2018). Some vendors allow the download of the data or offer a web *API* where data can be enquired.

Granularity of the data is usually decreasing the further away it is accessed from the source wearable. To decrease the burden on the network, data is summarised and aggregated (de Arriba-Pérez et al., 2016), hence, the degree of processing of the data increases. The time delay on when data is available also increases, since being transferred and processed in each layer takes time. The availability of the data also becomes less predictable the further away from the wearable source; data can just be transferred when a connection, i.e., Bluetooth, WiFi, or cellular internet, is available.

4.4. Discussion

As this chapter highlighted, there is no universal answer on how to obtain data from wearables in a way suitable for research (RQ 1). The four exemplary devices, which are representative for different data access approaches, showed very differing means of data granularity and accessibility.

Consumer vs. Professional devices. The level of data access differs for consumers of wearables and researchers. Naturally, when considering the *Data–Information–Knowledge–Wisdom (DIKW)* hierarchy (Section 4.1), consumers need highly preprocessed and contextualised views on data derived from sensors. In research context, access to lower level data is often required. In this context, the data access means for devices also differs. The professional devices, e.g., *Nexus 10 MK2* or Shimmer Sensors, naturally offer easy access to very fine-grained, low-level data, e.g., *Electrocardiogram (ECG)* signals — after all, that is what they are marketed for. On the contrary, the consumer devices offer less trivial access to numerical data, but they often offer summarised and visualised information to users. Numeric data usable for research is mainly accessible programatically via *Application Programming Interfaces (APIs)*/*Software Development Kits (SDKs)* or through third-party apps.

Data Granularity. Especially in consumer devices, which are purposed to provide informations in an easily to digest manner for casual users, the information derived from sensing data is often presented in form of summaries and easy-to-grasp charts (Karapanos et al., 2016). Often, these visualisations are shown either on the wearable itself, in a mobile companion app, or an online dashboard. For researchers to utilise the wearable sensing data, it not just has to be accessible in a research-appropriate, i.e, quantifiable form, but its level of granularity has to match the research scope. This means that a) sampling rates of the data have to be suitable for the proposed analysis and modelling, and b) the level of preprocessing has to be appropriate. While some research applications can utilise higher-level features derived from sensing data — providing that the data is valid and reliable — other research applications need access to raw, unprocessed data, e.g., raw accelerometer data for fall-detection.

Location of Access. The 'location' of the data access differs. Depending on the available access means by the wearable device manufacturer, data can be either accessible on the wearable, on the gateway device, or potential internet and cloud services. Access of data on the wearable itself can burden the researcher with the task to actually transmit the data to device where it then can be downloaded from. For example, the *AWSense* framework eases this task of data transmission and storage in a downloadable format for the *Apple Watch*.

Real-time vs. historic access. The temporal component of data access differs. Naturally, the further away the access point is the more time delay is introduced, i.e., accessing data streams on the wearable itself (e.g., *Apple Watch*) can occur with less time delay than access from a cloud server. The granularity and level of processing also matters; higher-processed data, which also is mostly found further away from the sensing device, i.e., the cloud, is usually just accessible as historic data.

Device specification, documentation, and standardisation. What became apparent from considering the 3 consumer devices within this chapter, is a difference in how sensor specifications are documented. While Microsoft (2016) provides a detailed documentation on the specification of the provided sensing data including sampling rates, Apple (2016) does not disclose the sampling rate for its heart rate sensors. This makes relying on the provided data difficult for researchers. A detailed specification and documentation of the provided sensing data is crucial for researchers to firstly understand what data is available and how it can be accessed. Additionally, it is crucial for researcher to know how the data has been preprocessed to make assumptions on its suitability. Lastly, standardisations for data access exist, e.g., the *Generic Attributes (GATT)* Bluetooth profile for providing heart rate data over *Bluetooth Low Energy (BLE)*. Devices supporting this standard are, e.g., Bluetooth heart rate straps or selected wrist devices such as the Mio devices.[17] This standard further ensures, that the format of the provided data is clearly defined. Altogether, this enables cross-interoperability and easier connection of devices. Further, new devices supporting these standards could easily be integrated in existing systems. More work is needed to expand these standardisation efforts and promote their implementation in commercial devices.

Ressource Efficiency. For some research setups, a resource efficient collection of sensing data may be necessary, e.g., a data collection app which drains the battery of study participants' device can be frustrating. This can lead to a lower compliance and increased drop-outs of participants. As exemplarily shown in Section 4.2.3, collecting *Apple Watch* sensing data with the *AWSense* framework has implications in the battery life of the device. Depending on the configurability of the wearable device, strategies for preserving resources can be considered, e.g., sensing data collection with lower sampling rates or just when it is needed. Further, thorough testing and benchmarking of a sensing application could be performed. Future research could explore strategies to more efficiently collect sensing data, e.g., as discussed for mobile sensing by Ben Abdesslem et al. (2009).

Collection of Subjective Data and Feedback. Sometimes, the collection of mere sensing data is not enough for research purposes but rather there is a need for subjective data to be collected from participants. In case of laboratory and on-site studies, this data can be collected by the researchers using, e.g., paper questionnaires or audio interview recordings. In case of remote studies, it could become necessary to either collect data on the participant's phone or even wearable. Modern, programmable smartwatches offer *user interface (UI)s* which can be used to collect additional and subjective data from users. Research has shown that using wrist wearables to collect experience samples from participants, led to a decreased response time compared to

[17] `https://support.mioglobal.com/hc/en-us/articles/229970728-What-other-apps-are-compatible-with-Mio-devices-` (accessed: 07/06/2019)

the mobile phone (Hernandez et al., 2016). Similarly, feedback, e.g., just-in-time interventions, can be provided on the wearable instead of the smartphone. Future work should explore if this leads to feedback and interventions being easier accessible for participants. Although the main focus of this thesis is on physiological sensing data, a programmable *UI* can be an important criteria for choosing a device depending on the research setting.

Privacy. Modern consumer wearables often have complex architectures and data flows, where collected data of the user is not just stored on the wearable or personal phone, but it is also transferred to cloud servers. Depending on the policies of the vendor, this data could potentially shared with third parties. Following, sensitive data of participants could be shared with these parties, e.g., health or behavioural data. Data ownership not just relates to the user's owning and being able to fully access and explore their own data, but also the rights of what vendors can do with the user's data (Ball, 2015). Santos (2016) explains that device manufacturers often claim rights on the users' data within their privacy policies and user agreements; these rights allow the sharing and usage of data for researching trends or marketing purposes. Segura Anaya et al. (2018) found that users of wearable devices consider informed consent on wearable data usage by third parties as 'very important.' While their survey considered personal device usage, the same applies to research settings. It should be the researchers responsibility to examine data policies of the device vendor, consider ethical and privacy concerns, and clearly communicate any implications to participants for these to provide informed consent.

4.4.1. Data Access Implications for using Wearable Devices in Research

Choosing the a wearable device for a research setup which provides the right kind of data in a suitable way is not trivial. When considering a specific wearable device, aspects such as the types of sensors and the nature of the provided data have to be considered.

4.4.1.1. Set of sensors and granularity

Naturally, the types of sensors provided must match the research scope and hypotheses of a project (e.g., research on the heart activity of patients should include sensors to provide appropriate data on the subjects heart, such as *Photoplethysmography (PPG)* or *ECG* sensors). Beyond the availability of the right 'kind' of sensors in a device, also the right granularity need to be considered.

Especially consumer devices often do not offer fine-grained data access or even access to raw data signals, e.g., the *PPG* signals[18]. The questions researchers need to consider are:

a) Does a device contain the right sensor for the research problem? — e.g., is a heart rate sensor included

b) Does the device manufacturer allow access to the right kind of data from a sensor? — e.g., access to *RR-Interval* from a heart rate sensor or access to the raw *PPG* signal for applying own signal processing.

c) Is the provided data fine-grained enough for the research problem? — e.g., the sampling rate and precision is appropriate

d) In case that subjective user data is needed, is it possible and feasible to use the wearable's *UI* for this? — e.g., for the administration of several short questionnaires directly on the wearable

4.4.1.2. Data Access Means

The ways of accessing device data vary amongst the devices. Different architectures and approaches have been implemented in wearable systems. Depending on the nature of the research where a device should be applied, devices can have a varying degree of suitability.

Especially consumer devices are often designed with the end user in mind who may not require any access to the collected sensing data beyond visualisations or summaries. Out-of-the-box support to obtain research-appropriate data (e.g., in the right format or granularity) are scarce. In some cases, third-party developers offer apps for data access utilising the manufacturer provided *APIs* or *SDKs*. An example is *Heart Rate Variability Logger*[19] for iPhones by Altini (2013); this app not just enables the download of the collected sensing data from selected heart rate devices (e.g., *Polar H7*) but it also calculates additional *Heart Rate Variability (HRV)* features. While such third-party apps could be an option when a device is used in a laboratory setting or localised studies, e.g., Gabana et al. (2017), they are mostly unsuitable for remote in-the-wild studies where access to the data needs to be performed from afar.

[18] An overview of the shape of the raw signal obtained from a *PPG* sensor can be found in Section 2.2.2.1
[19] `https://itunes.apple.com/gb/app/heart-rate-variability-logger/id683984776` (accessed: 07/05/2018)

Alternatively, the development of own data collection applications can be an option;[20] this depends on the available resources of the researcher and the manufacturer's provided *APIs* or *SDKs*. In case that the data collection is planned to happen on a larger scale and by participants using their own wearable devices, future developments and the fast pace of the marked need to be considered. Relying on devices with high market shares allows to reach a higher number of potential participants. Similarly, relying on devices with standardised protocols can make the support of a variety of current and future devices easier.

Lastly, and no matter whether third-party applications or custom data collections means are utilised, ethical and privacy-preserving handling of the participants' data needs to be adhered. This is even more so relevant when potentially very sensitive physiological data is collected. Privacy policies of the device manufacturers should be considered. The ease of capturing data with digital devices like wearables also can tempt researchers to tap into any data stream they could get hold of without actual necessity; this should be prevented (Mok et al., 2015).

Depending on the research scope, it has to be considered if a real-time data access is necessary. This may, for example, be the case if the data is instantly processed and actions, such as just-in-time interventions, are taken. In case that the data is merely analysed at a later stage, it is sufficient to obtain the historic data, e.g., from an online data warehouse.

The questions researchers need to consider are:

a) Are the access means appropriate for the researcher's available resources? — e.g., device data can just be accessed via expensive third-party applications, or device data can be just accessed via customary implemented code (e.g., *Polar H7*)

b) In case of a more complex research pipeline or scenarios, are the access means compatible and integratable? — e.g., wearable device data needs to be integrated in an online cloud system, so a web *API* (e.g., *Microsoft Band* Health cloud) is favourable

c) Depending on the research scope, is real-time or time-delayed access appropriate? — e.g., data-driven just-in-time interventions require real-time data access.

[20]For the laboratory experiment in Section 5.2, a custom app to collect data from the *Apple Watch*, *Microsoft Band 2*, and *Polar H7* was implemented

d) Is the use of a certain device and its manufacturers data policies applicable to the study in terms of ethical and privacy concerns? — e.g., sensitive data of participants is stored on a could service and shared with third-parties

e) In case of participants using their own device, is the device common so enough users can be recruited for a study? Is it possible to expand the pool of potential users by relying on widely supported standards? — e.g., the *BLE* heart rate profile is supported by multiple vendors and devices

4.4.2. Summary

The practical consideration of four state-of-the-art devices and related literature has highlighted that there are important factors regarding the availability and granularity of sensing data accessible from wearable devices. This boils down to two factor to consider: a) the granularity and richness of the data, and b) the effort and means needed to access the data in a research-suitable manner. Those two factors in themselves are complex and not easy to answer, even for a given device. Depending on the research question, scope, and design, the researcher should consider the above proposed questions to evaluate the suitability of a given wearable device. The research question (RQ 1 — *How can research-suitable data be obtained from wearable devices?*) cannot be universally answered; but rather there is a set of guidelines and considerations for researchers to take into account.

The outcomes from this chapter are utilised in the following ways:

a) The apps and frameworks presented in this study were utilised in Chapter 5 on evaluating the presented devices in a laboratory and in-the-wild setup for their ability to detect stress and affect.

b) The lessons learned in this chapter will be utilised to develop an *Initial Design Space (IDS)* on criteria for choosing wearable measurement devices in research setups (Chapter 6).

Chapter 5

Research Probes: Stress and Affect in the Lab and the Wild

To explore how suitable 4 state-of-the-art wearable sensing devices are to measure stress and affect, two studies were conducted. The first study presents result from utilising the Apple Watch to collect sensing data and subjective mood experience samples in everyday life. Secondly, a laboratory experiment evaluated 4 devices in stressful and physically active situations. It was shown that several shortcomings with these devices exist. Finally, a discussion of the results, experiences and lessons is presented.

Wearable devices with physiological sensors incorporated show potential to provide insights in complex internal processes of the within the human body. Sensors such as heart rate, *Electrodermal Activity (EDA)* or skin temperature have been widely applied to detect stress and affect[1]. The previously discussed test devices — namely *Apple Watch*, *Polar H7*, *Microsoft Band 2*, and *Nexus 10 MK2* — each include a subset of these physiological sensors (see Section 4.2). But the devices suitability to provide valid sensing data to detect stress and affect states has not been previously evaluated. Consequently, this chapter focuses on RQ 2: *How suitable are current state-of-the-art wearable devices to be applied for measuring stress and affect in the lab and-in-the wild?*

To address this question, two research probes are presented. Firstly, the *Apple Watch* smartwatch is examined for its ability to show a relationship between sensing data and affect mood samples in the wild. Secondly, the four test devices are evaluated for their ability to detect stress and

[1] see Background Section 2.3

affect in the lab. Based on these studies, the suitability for affect and stress detection of the four exemplary devices is discussed. Further, it is summarised how researchers should proceed to evaluate novel and previously untested devices. Finally, the findings of these studies informed the conceptualisation of the *Initial Design Space (IDS)* of the following Chapter 6.

5.1. Using the Apple Watch for affective research in the wild

The *Apple Watch* is a popular smartwatch with a high market share; this makes the device attractive for large-scale, in-the-wild research studies. The options to deploy apps for data collections and interventions through the App Store, allows for research studies to be conducted remotely and without the need of participants to be local. The *Apple Watch* is particularly designed to integrate with Apple's health platform for data storage of summarised fitness metrics. *HealthKit* allows developers and researchers access to this stored health metrics[2]. Apple (2015), furthermore, provides the open source framework *ResearchKit* for iOS apps. *ResearchKit* eases the creation of research apps through providing templates for consent forms, mechanisms for experiments, tools to create assessments on the phone. It makes remotely-controlled research on the iPhone more accessible. Currently, there are apps using *ResearchKit* for Asthma, Diabetes, and cardiovascular health (Icahn School of Medicine at Mount Sinai, 2015; Zhu, 2015). In summary, this makes the *Apple Watch* attractive for research settings.

The objective of this study was to evaluate sensing and affective data collection from the *Apple Watch* in the wild. As discussed before, the *Apple Watch* is equipped with a *Photoplethysmography (PPG)* heart rate sensor which has the potential to provide data related to the affective mood state of the wearer. To test the suitability of the *Apple Watch* for distributed studies, the below study was designed to be deployed to participants remotely and through minimal interference of the researcher.

[2] A detailed overview of the data access from the *Apple Watch* is discussed in section 4.2

Within this research probe, an *Apple Watch* app collects wearable sensing data and affective data through experience sampling during participants everyday life. The *Experience Sampling Method (ESM)* by Larson and Csikszentmihalyi (1983) is a research method where participants are asked to note their right-now experiences, like feelings, thoughts and the current situation. The goal is to obtain a representative sample of experiences from a user in a natural environment and outside of controlled lab conditions. Within this study, *ESM* is used to obtain data about the user's affective states at various points during the day. These affective experience samples are collected directly from the wrist of the participant by using the *Apple Watch*'s *user interface (UI)*. Work by Hernandez et al. (2016) has shown that collecting experience samples on the wrist opposed to the phone leads to decreased response times. This subjective data is enriched by objective sensing data from the *Apple Watch*, like heart rate, physical activity and wrist movement data. The *AWSense* framework[3] is used for collecting the sensing samples.

5.1.1. Study Aim and Hypotheses

The aim was to explore how suitable this specific device is a) in terms of collecting sensing data remotely, and b) in terms of the suitability of the data to show a relationship between sensing and subjective, sampled mood. Related work has shown that neurophysiological processes of the *Autonomous Nervous System (ANS)*, which are measurable through changes in physiological signals, are related to subjectively perceived mood[4]. Instead of looking into distinct mood categories, e.g., happiness, anger, etc., a two-dimensional affect model following Russell (1980) has been considered in this study[5]. This model describes two dimensions of affect: pleasantness (valence) and activation (arousal). Related work has shown that heart rate increased when people experienced higher arousal and higher valence (Witvliet and Vrana, 2007; Salimpoor et al., 2009). Regarding the relationship between subjective and physiological measures, the following sub-hypotheses are derived:

Hypothesis 5-1a *There is a correlation between arousal and heart rate data from the Apple Watch.*

Hypothesis 5-1b *There is a correlation between valence and heart rate data from the Apple Watch*

[3]See Section 4.2.3 for information on the *AWSense* framework
[4]See background Section 2.3 for physiological signals and affect
[5]See background Section 2.3.1 for dimensional affect models

As commonly acknowledged, physical activity has an impact on physiological processes within the body. For example, physical activity is related to an increase in heart rate (Boettger et al., 2010). Therefore, it is expected to see an impact of physical activity, namely step count, just before a heart rate sample is taken:

Hypothesis 5-2 *There is a relationship between step count and heart rate from the Apple Watch.*

Further, it is explored how participants felt about using the provided *Apple Watch* app for logging their affective state during their everyday life.

5.1.2. Study Design and Material

Since the aim of the study is to explore the *Apple Watch*'s suitability to show a relationship of sensing data with subjectively collected mood samples, a non-experimental design is chosen. The relevant measured data is collected without imposing experimental conditions on the participants.

5.1.2.1. Measured Variables

With the main Hypotheses 5-1a and 5-1b to explore, firstly, affect experience samples in form of valence and arousal are collected from participants, and secondly, heart rate sensing data is collected. Further and with respect to Hypothesis 5-2, physical activity features such as motion features and secondary health features, i.e., step count are recorded.

5.1.2.2. EmoRate App

To conduct the proposed study and collect sensing and subjective data from the *Apple Watch*, the *EmoRate* app has been developed. In summary, the *EmoRate* app is an app for the iPhone and *Apple Watch* where mood experience samples in form of valence and arousal ratings are collected using the wearable *UI*. Additionally, sensing data in form of heart rate and physical activity is collected. The app is designed to work as a standalone application without the need of the research team interfering; thus, it can be used in remote studies. All the important components of the study, such as informing participants, obtaining consent, and secure and privacy conserving data collection have been integrated in the app. A detailed description of the app and its functionalities is presented in Appendix C.

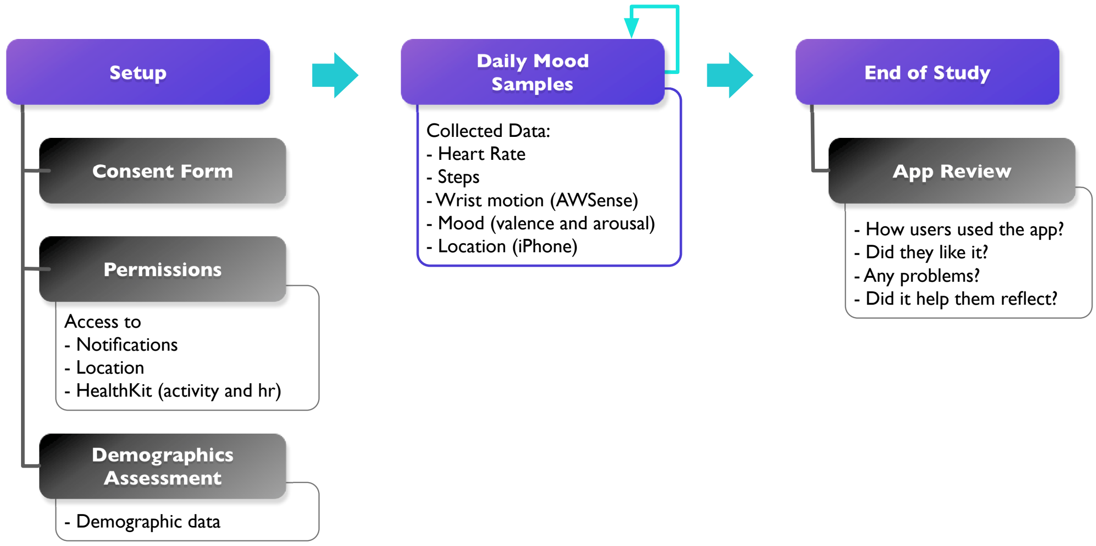


Figure 5.1.: Study process of the EmoRate pilot study. Depicted it the setup process with the informed consent, sensor access permission, and initial assessment of the participants, the process of collecting daily mood samples, and the final assessment.

5.1.2.3. Study Procedure

Participants for the study were recruited using mailing lists. They were required to own an *Apple Watch* to ensure familiarity with the device. Further, the participants would use their own devices during the study. The *EmoRate* app was distributed to the participants using Apple's *TestFlight*. This allowed invited participants to install the app on their personal phone and *Apple Watch*. Participants were briefed, that they could start using the app whenever they liked and that the study would last seven days. From there on, everything was managed from within the *EmoRate* app. An overview of the study process and collected information is shown in Figure 5.1.

During the setup process, users were provided with information on the study and informed consent sheet to be signed on the phone. The signed PDF version was sent to the researcher. An exemplary, PDF version of the consent form (signed by the Thesis author), as well as, the ethical approval letter can be found in Appendix D. After signing the consent form, participants were asked to provide demographic information, i.e., age, gender, and occupation. The survey was designed with *ResearchKit*.

ID	age range	gender	occupation	no of assessments
P1	18-25	female	student	37
P2	36-45	male	professional	38
P3	36-45	male	professional	21
P4	26-35	female	student	33
P5	26-35	male	professional	20
P6	26-35	female	professional	44

Table 5.1.: Demographic profile of participants in the *EmoRate* pilot study and the number of mood assessments they provided

Following this setup, participants were promoted to provide emotional experience samples six times per day[6]. When participants reacted to this notification, the *EmoRate* app opened on the watch and participants were able to rate their affect using a 5-point arousal and valence scale. The valence and arousal icons for presenting the different options are loosely adapted from the *Self-Assessment Manikin* scale by Bradley and Lang (1994); the *UI* is shown in Appendix C. When the app was open, sensing data from the *Apple Watch* was collected in form of heart rate and device motion data. This data and the historical step count data 30 minutes before the assessment were exported for later analysis.

On the final day of the study, participants were asked to fill out a questionnaire on their phone. It focused on their experience with the app and improvements.

5.1.2.4. Participant Profile

The *EmoRate* app was administered in a small pilot study with 6 participants; all of the participants finished the 7-day study. Participants were recruited via word of mouth. Within this study, a total of 193 emotional assessment samples were collected; an overview of the demographics and the number of mood assessments per participant can be seen in Table 5.1.

[6]The reminders were scheduled at 9:00am, 11:00am, 1:00pm, 4:00pm, 6:00pm, 8:00pm

5.1.3. Results

The *EmoRate* app was administered in a small pilot study with 6 participants; all of the participants finished the 7-day study. Within this pilot, a total of 193 emotional assessment samples were collected; an overview of the demographics and the number of daily assessments per participant can be seen in Table 5.1.

The limitations of this pilot study lay clearly in the small number of participants and the short duration. Still, some findings and an overview of the collected data will be given below.

The main contribution lay in evaluating in-the-wild sensing with the *Apple Watch* smartwatch and feedback on the app for collecting experience samples throughout the day on the wrist device.

Time dependence of the data. There are clear trends visible when considering the time dependence, i.e., the day of the week or hour of the day. The arousal and valence ratings both indicate an increase throughout the day meaning that participants experienced higher arousal and valence in the evening (see Figure 5.2). For arousal, this relationship was significant with a Spearman correlation coefficient of $\rho_S = .325, p = .000$. The valence showed a weak correlation with the time of the day; $\rho_S = .156, p = .034$).

Mean valence and arousal readings throughout the week were fairly consistent; valence peaked on a Saturday but decreased again on Sunday (cf., Figure 5.5). The average heart rate during the assessments increased from Monday to Thursday but was lower on Fridays and the weekend.

Relationship of subjective and sensing measures. A comparison of the subjectively assessed mood ratings (valence and arousal) with the collected sensing data (heart rate and wrist acceleration) revealed no significant correlations apart from a weak correlation of heart rate and valence (both intra-subject z-standardised to account for variation between the subjects physiology (c.f., Lykken et al. (1966)), e.g., resting heart rate); $\rho_P = .156, p = .037$ (see Figure 5.4). Especially a positive relationship between perceived arousal and heart rate was expected, as per Witvliet and Vrana (2007) and Quesada et al. (2014); this relationship could not be confirmed in the collected data from the pilot study.

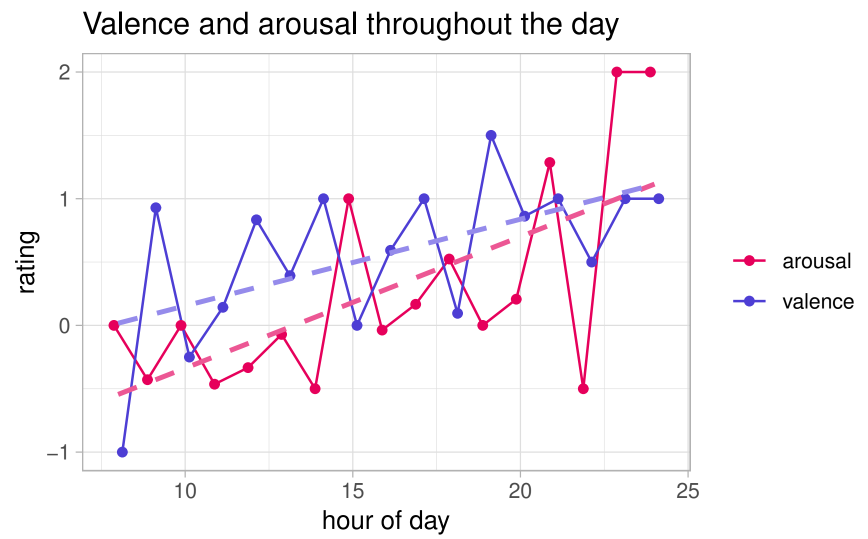


Figure 5.2.: Arousal and valence ratings throughout the day. The graph shows a tendency for an increased arousal and valence later in the day.

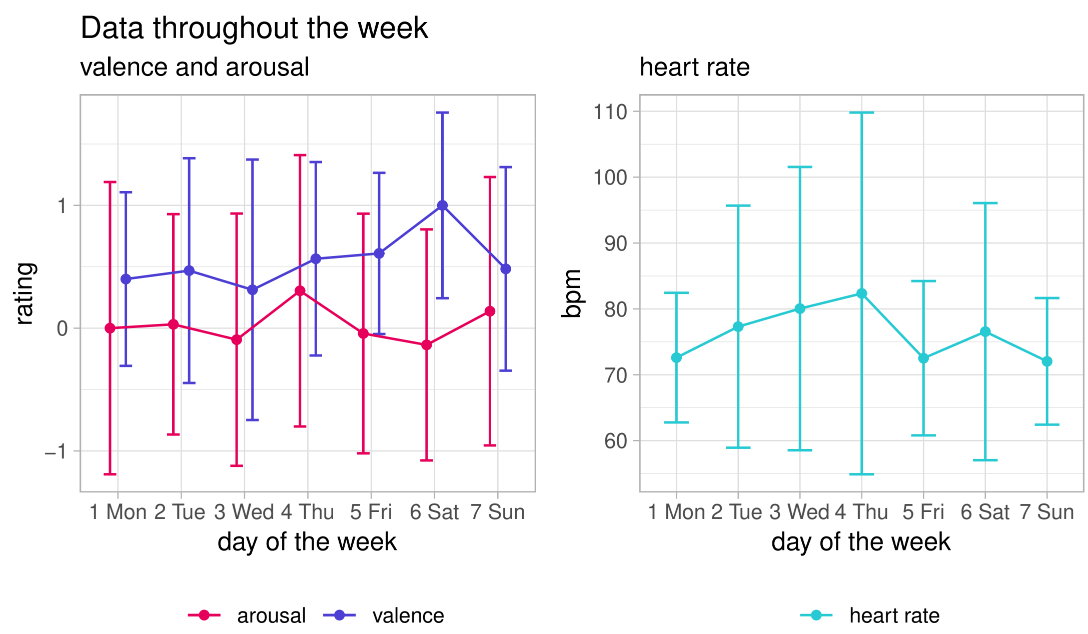


Figure 5.3.: Arousal, valence and heart rate throughout the week. The graph shows a tendency for higher arousal/valence and lower heart rate later in the week and on weekends.

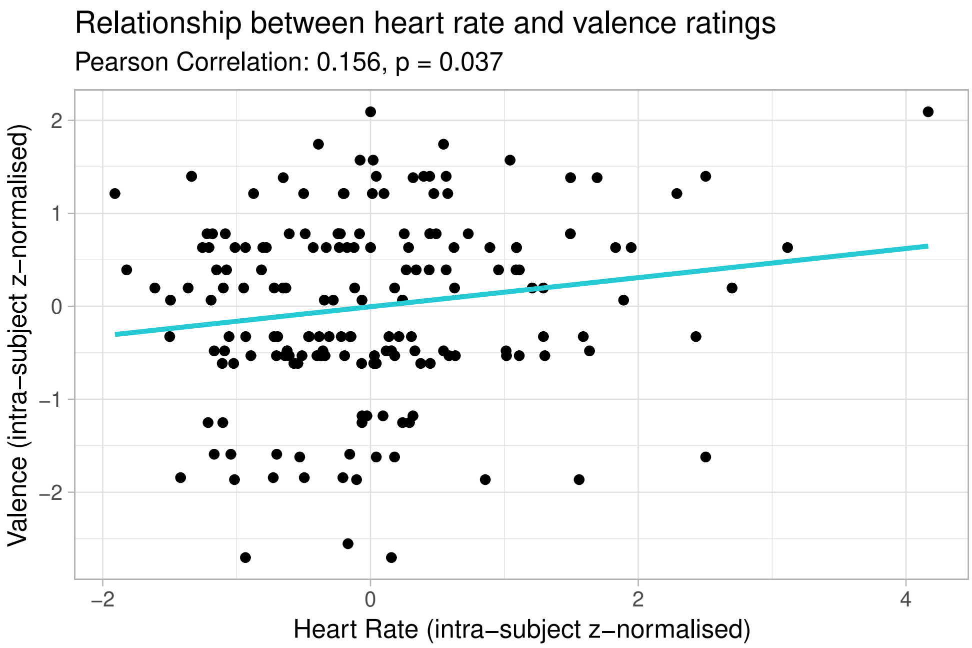


Figure 5.4.: Relationship of heart rate and valence ratings show a weak positive relationship ($\rho_P = .156, p = .037$). Both readings have been intra-subject z-standardised to account for variation in baseline readings between subjects.

Relationship of heart rate and movement data. Considering the relationship amongst the sensor readings, there was a positive relationship between heart rate readings during assessments and the wrist movement activity based on accelerometer readings. The strength of the wrist movements was calculated from the x, y, z components of the accelerometer vector **a** by calculating the magnitude (length) $|\mathbf{a}|$:

$$|\mathbf{a}| = \sqrt{\mathbf{a}_x^2 + \mathbf{a}_y^2 + \mathbf{a}_z^2}$$

The heart rate correlated significantly with both the mean accelerometer magnitude ($\rho_P = .337, p = .000$) and the variation (standard deviation) in the magnitude ($\rho_P = .246, p = .001$). Possible explanations could be higher physical activity leading to both higher heart rate and wrist movement.

Very naturally, there was also a positive relationship between the heart rate readings and the number of steps 10 minutes before each assessment. Considering absolute heart rate readings,

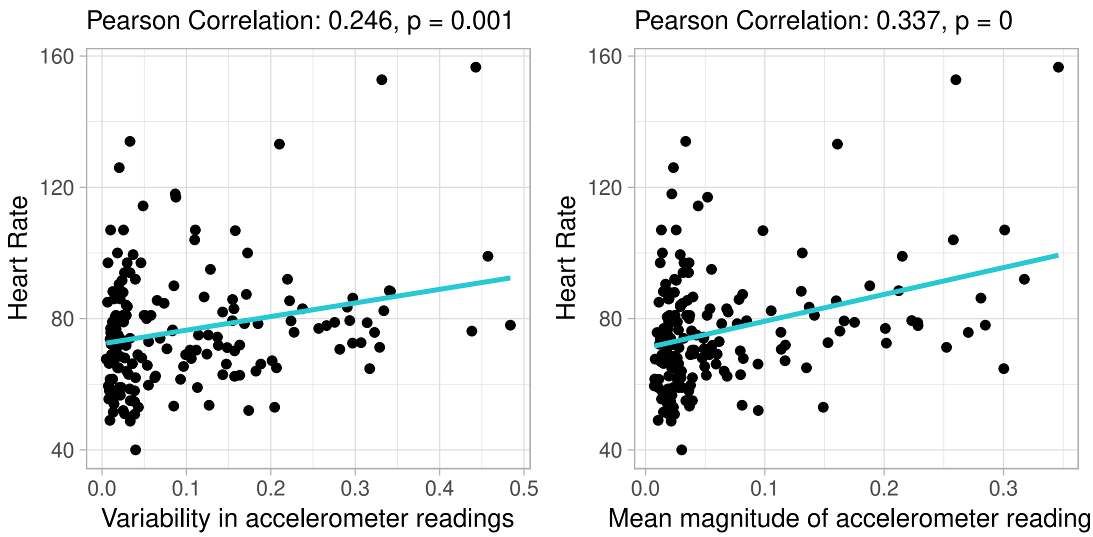


Figure 5.5.: Relationship between average heart rate during assessments and the wrist movements. There are significant positive correlations between heart rate and mean wrist acceleration and variation (standard deviation) of wrist acceleration.

there was a correlation of $\rho_P = .297, p = .010$. Considering the z-normalised heart rate readings to account for inter-subject variation, the correlation increased to $\rho_P = .354, p = .002$. The relationship between steps taken and heart rate is not novel, e.g., (Lubans et al., 2009).

User feedback on the app. At the end of the study, user feedback on how participants experienced and used the app was collected.

The first part of the questionnaire focused on the satisfaction with the app and the ease of use. The majority of participants were satisfied with the app. Furthermore, all participants perceived the app very easy to use and rated 5 or above on the 10-*Likert Scale* (cf., Figure 5.6). Two participants suggested improvements in the form of *'In app [sic] instructions on how to use the phone app when you first open it'* (P5) and *'[i]nstruction[s] on using [the] crown to make selection'* (P2).

Four out of six participants also indicated a tendency to keep using the app in the future. The daily reminders were also perceived as useful. Most participants also used the phone app to review their emotional ratings. Participant stated:

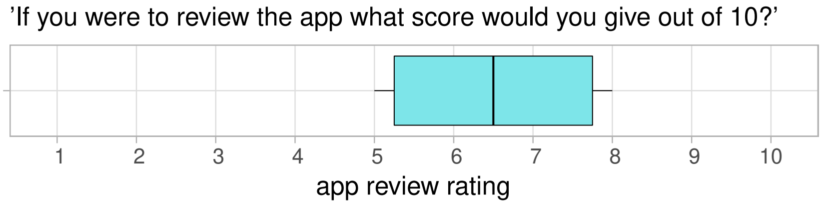


Figure 5.6.: Final App Review of the *EmoRate* app

- *'Was cool but would like a bit more feedback about my mood maybe'* (P4)
- *'More feedback on what I input'* (P4)
- *'It will more helpful to review moods if people could visualise their records in a daily or weekly basis'* (P1)

The collected user opinion indicates the need for more feedback and better visualisations to enhance the app and give the app a purpose beyond data collection of keeping participants engaged.

5.1.4. Summary

The *Apple Watch* as a popular smartwatch with integrated sensors (i.e., heart rate, accelerometer, gyroscope, magnetometer, *Global Positioning System (GPS)*) and a simple programmable *UI* poses as a gripping device for researchers to be used in human-subject research. Larger medical studies already leverage the *Apple Watch* for remote data collection in the wild, e.g., EpiWatch seizure tracking by John Hopkins Medicine (2018), the Apple Heart Study by Stanford Medicine (2018), or heart rate tracking during concussions by NYU Langone Health (2018). But the consideration of the *Apple Watch* for affect sensing in the wild is novel. Within this study, the feasibility to use the *Apple Watch* to collect sensing data and low-intrusive affect experience samples from the device itself has been demonstrated. Participants were satisfied with the *EmoRate* app for logging their emotions throughout the day; a subset of participants voiced even that they would keep using the app and wished for more insights on their collected data.

Furthermore, positive relationships between subjectively perceived valence and heart rate sensor readings were observed confirming Hypothesis 5-1b. Contrary to the expectation, no relationship was observed for arousal; Hypothesis 5-1a could not be confirmed. Naturally and as secondary findings, an increase in the physical activity and step count was also related to an increased heart rate. These findings highlight that both physical activity and emotional state influence heart rate, and this should be taken into account when interpreting the cause for heart rate fluctuations. The limitations of this pilot study lay clearly in the small number of participants and the short duration.

The next section goes into details and evaluates three physiological sensing wearable devices in a laboratory setting under physical activity and stress-inducing tasks.

5.2. Evaluation of Wearable Sensing Technologies to Detect Stress in the Lab

While there is an evermore flood of wearable gadgets entering the market, little work is done on actually validating the promises and data of these devices. The sheer amount of devices makes it not feasible to evaluate all of them and there is no unified procedure of manufacturers to prove reliability or lay open internal data processing and smoothing mechanisms. Peer-reviewed evaluations, mainly from the fields of sports science and medicine, focus on the reliability of mainly fitness related measures, such as step counts, energy expenditure, or heart rate readings (Evenson et al., 2015; El-Amrawy and Nounou, 2015). But novel wearables promise to even detect sensitive states, such as stress[7] by using physiological signals from fitness trackers.

5.2.1. Study Aim

The aim of this laboratory experiment is to validate the exemplary wearable sensing devices for their ability to infer stress and increased arousal. The devices considered, are the same devices

[7] See Garmin VivoSmart `www.garmin.com` (accessed: 20/05/2018)

discussed regarding the data access in the previous Chapter. These devices offer relevant physiological sensors often used in affective computing and stress detection, namely heart rate, skin conductance, and skin temperature[8]. These devices shall be evaluated within a controlled and hypothesis-led experiment layout.

5.2.2. Hypotheses

Within this experiment, the devices' ability to identify differences in sensing data during low and high physiological arousal tasks under the consideration of physical activity effects. In the following, the reasoning and hypotheses are presented.

5.2.2.1. The Impact of Stress

Stress and undergoing a stress-inducing task has shown to have an impact on the subjectively perceived, but also sensory measurable physiological signals. The following sections draw the assumptions in terms of these effects.

Differences in Subjective Stress Measures under Low and High Stress. Firstly, it is assumed that subjectively perceived measures, such as perceived stress and arousal, are increased after performing a stressful task. This leads to the following hypothesis:

> **Hypothesis 5-3** *There is a difference in subjectively perceived stress measures (namely arousal and perceived stress) between low and high stress conditions.*

Differences in Physiological Stress Measures under Low and High Stress. As related work has shown and as discussed in Section 2.3, physiological signals and bio-markers differ under low and high arousal and stress. Arousal and stress responses are classified by a range of physiological changes and chain reactions, such as the release of hormones triggering an increase in heart activity, sweating or muscle tension. The measurable stress responses are characterised by an increased heart rate and skin conductance (*Electrodermal Activity (EDA)*), as well as, an decreased surface skin temperature. But previous findings show a decreased reliability of physiological sensing devices during physical activity (see Hypothesis 5-5a below). Therefore,

[8]For a detailed overview on the sensors, please refer to Section 4.2. A discussion on related work on using physiological sensing in affective computing can be found in Section 2.3

the assumption that sensing data differs during low and high mental stress tasks is considered separately in two sub-hypotheses:

> **Hypothesis 5-4a** *There is a difference in physiological data of each device between low and high stress situation under stationary activity.*
>
> **Hypothesis 5-4b** *There is a difference in physiological data of each device between low and high stress conditions under physical activity.*

The hypothised changed are an increase in heart rate, increase in skin temperature and decrease in skin temperature under the stress task.

5.2.2.2. The Impact of Physical Activity

Physical activity in itself has an impact on physiological processes within our body. Additionally, movement through activity can also disrupt sensor readings. The following paragraphs draw the assumptions based on different physical activity levels.

Differences Across Devices under Low and High Physical Activity. Physical activity and movement has an impact on readings of body-placed sensors. Movements and vibrations can cause slight slipping of the sensor on the skin which in turn can lead to undesired noise and artefacts in the data. Especially the *Photoplethysmography (PPG)* signal, like it is used in a majority of modern fitness tracking watches, has been shown to be prone to movement artefacts (Allen, 2007). Validation studies, e.g., by Tamura et al. (2014), showed a decreased accuracy of wrist-worn measured heart rate in consumer devices.

Furthermore, studies have shown an increased error percentage in sensor readings while participants are moving (Shcherbina et al., 2017). Based on these findings, the following hypotheses are derives:

> **Hypothesis 5-5a** *There is a difference in the physiological data measured by different devices under physical activity*
>
> **Hypothesis 5-5b** *There is no difference in the physiological data measured by different devices under stationary activity*

Hypothesis 5-5c *There is a significant increase in error percentage compared to the reference device under physical activity.*

Changes in Physiological Measures under Low and High Physical Activity. Opposed to mental stress and psychological stress, physical activity can be seen as a stressor in itself with similar reactions, namely increased heart rate and skin conductance (*EDA*), as well as, a decreased surface skin temperature (Boettger et al., 2010). Therefore, the following hypotheses were derived:

Hypothesis 5-6a *There is a difference in low and high physical activity sensor readings and subjective data in relaxed conditions.*

Hypothesis 5-6b *There is a difference in low and high physical activity sensor readings and subjective data in mentally-stressful conditions.*

5.2.2.3. Relationship between Subjective and Physiological Measures

Lastly, there is a hypothesised relationship between physiological sensing data and subjectively perceived measures. Common affect measures are hereby subjective arousal, valence, and perceived stress. The measures of arousal, valence and dominance have been assessed using the *Self-Assessment Manikin* (Bradley and Lang, 1994). Schimmack and Grob (2000) considered the measure of arousal to broad and argued that *wakefulness* and *tension* are independent sub-categories.

Prior work from neuropsychology suggested that there is a relationship between neurobiological processes and subjectively perceived stress (Gianaros et al., 2012; Schaefer et al., 2014), arousal (Gianaros et al., 2012; Quesada et al., 2014), and valence (Gianaros et al., 2012; Quesada et al., 2014). Other studies discovered that heart rate activity increased when subjective arousal and valence were higher (Witvliet and Vrana, 2007). Salimpoor et al. (2009) showed that arousal and valence strongly correlated with electrodermal activity, body temperature, heart and respiration rate as well as blood volume pulse. On the contrary, dominance was not found to be correlating with an increase of stress hormones (Quesada et al., 2014). Regarding the relationship between subjective and physiological measures, the following sub-hypotheses are derived:

Hypothesis 5-7a *There is a correlation in between stress perception and physiological data*

Hypothesis 5-7b *There is a correlation between arousal (wake and tense) and physiological data*

Hypothesis 5-7c *There is a correlation between valence and physiological data*

Hypothesis 5-7d *There is no correlation between dominance and physiological data*

Considering these hypotheses are the foundation for the choice of suitable stress-indicating measures, appropriate test devices and the design of a controlled laboratory experiment.

5.2.3. Study Design

As identified in the last section, three different consumer wearable devices and a laboratory reference device were tested for their suitability to detect changes in sensor data under different physiological arousal states induced by mental stress. Additionally, the impact of physical activity and movement on sensor readings was investigated. The two independent variables for the study were *physical activity* and *mental stress*. Both independent variables consisted of two levels each and a factorial design led to four overall experimental conditions, depicted in Table 5.2.

The overall experimental approach followed a within-subject design where every participant underwent all experimental conditions. This approach was chosen due to a high inter-subject variation of physiological signals, e.g., resting heart rate. To avoid ordering effects, the sequence of the conditions has been counterbalanced.

5.2.3.1. Independent Variables (IV)

The main two *independent variables (IVs)* in this study were *mental stress* and *physical activity*. These *IVs* informed the setup with the 4 experimental conditions. Furthermore, the four sensing devices were treated as *IVs* in terms of conforming the study hypotheses.

Mental Stress. To test the Hypotheses 5-4a and 5-4b that the sensing devices can detect changes in physiological signals induced by stress, the amount of mental stress put on the participants was varied. This *IV* was represented by two levels: a low mental stress level where people listened to meditation music and a high mental stress level where participants performed *Mental Arithmetic Tasks (MAT)s*; a task shown to increase perceived stress and activity of the *Sympathetic*

		Physical Activity	
		Low (**S**tationary)	High (**W**alking)
Mental Stress	Low (**R**elaxing Music)	**RS** (Baseline)	**RW** (3rd task)
	High (**M**ental Arithmetic Task)	**MS** (2nd/4th task)	**MW** (2nd/4th task)

Table 5.2.: Factorial design of experimental conditions based on the two independent variables *mental stress* and *physical activity*; resulting in four conditions: abbrev. RS, RW, MS, MW. The order of the non-stressful tasks were fixed. The order of the stressful tasks was randomised per participant.

Nervous System (SNS) causing physiological, measurable stress responses (Callister et al., 1992).

Physical Activity. As assumed by Hypothesis 5-5a, movement has an impact on the sensor readings due to increased noise in the data compared to consistent readings across the devices when there is no movement (Hypothesis 5-5b). To test this, participants were asked to sit stationary on a chair in the low physical activity level of this *IV* and walk on a treadmill in the high physical activity level.

Sensing Devices. In hindsight of the objective to validate distinct wearable devices with different sensing technologies,[9] these devices acted as another *IV*. The devices testes were: two wirst-worn consumer devices with *PPG* heart rate technology (*Apple Watch* and *Microsoft Band 2*), a consumer chest strap with *Electrocardiogram (ECG)* technology, and a professional biofeedback sensing device with *ECG* self-adhesive electrodes (*Nexus 10 MK2*).

[9]The wearables' sensing capabilities have been discussed in Section 4.2

5.2.3.2. Dependant Variables (DV)

Aiming to detect differences in the participants' physiological data, theses values were recorded using each of the four chosen measurement devices. Moreover, the focus was on potentially observable discrepancies between the physiological data suggesting stress and the self-reported arousal, valence, dominance, and stress ratings.

Physiological Data. From the aforementioned devices – *Nexus*, *Apple Watch*, *Microsoft Band*, and *Polar H7* – physiological data in form of heart rate, skin conductance (*EDA*), and skin temperature were recorded.[10]

Subjective Data. To assess the subjectively perceived stress and current affect state, the *SAM* and three scales on assessing tense, wake arousal and perceived stress were used.[11]

5.2.3.3. Stimulus Material and Tasks

Our stimulus material consisted of *MATs* which had been proven to elicit stress in participants (Biederman, 1973) and also influence physiological measures (Hassellund et al., 2010; Linden, 1991; Seraganian et al., 1997; Tomaka et al., 1994).

The *MATs* were presented on a a 60-inch display being positioned on a portable stand in front of our participants. The display was connected to the experimenter's laptop and showed the mathematical exercises. Figure A.11b shows the exercise screen as it is seen by the participants; it shows the exercise and a time bar. The time bar was added as an additional stressors in form of time pressure. Setz et al. (2010) showed that this is a valid method to further increase stress. For each of the calculations, addition and subtraction of two-digit numbers ranging from 0-100 and including negative solutions, the participants were given six seconds each. They had to speak the answer out loud. The correctness of the answer was logged by the experimenter, who saw the answers on her screen, by pressing the according button for a wrong or right solution[12]; this resulted in the participant receiving instant feedback on their screen. While for correct answers the screen turned green and the word "Correct" was displayed, false answers or time

[10] These devices and their provided sensing data are extensively discussed in Section 4.2.

[11] An overview of affect models can be found in Section 2.3.1.

[12] The experimenter screen contained the solution in words to make it easier for the experimenter of German origin, where two-digit numbers are spoken in reverse, to compare the spoken solution of the participant and the actual solution.

outs were signified with either "False" or "Time out" being shown on a red screen and a buzz sound was played. The tool to present the exercises, time bar and feedback was developed by the author and is open-source available [13]; screenshots can be found in Appendix F.

According to our study design, the relaxing task consisted of listening to meditation music[14]. The study setup was inspired by Vlemincx et al. (2012). Regarding the performance of physical activity, we asked our participants to walk for five minutes on a treadmill (model: ProFitness Sierra motorised) for the walking task.

5.2.3.4. Participants and Procedure

In total, 24 participants were acquired via university mailing lists, leaflets and personal recruitment campaigns. For the final data analysis, 3 participants were excluded: one pilot participant who had volunteered for a test trial of the laboratory setup and two other participants due to technical problems during the data collection. Hence, 21 participants with a mean age of 28.9 years (SD = 4.5) remained; among them were 8 females and 13 males. The recruitment process included strict exclusion criteria, namely no participants being taken who had been diagnosed with any heart conditions, mental illnesses or learning disabilities. Participants also had to guarantee that they did not suffer from alcohol and/or drug addiction. Before the experiment started, they were asked to refrain three hours from caffeine. For compensation each participant received 15£.

Prior to the study, participants were introduced to the experiment environment at Body-Centric Lab of the Queen Mary University London (Figure 5.7). The Body-Centric Lab is a quiet, undisturbed space which is protected from environmental temperature and light changes.

Participants were briefed about the study background as well as the sensor placement on the body. Following, they were asked to sign the consent form and to fill in an initial assessment consisting of demographic questions, self-reported fitness level, and smoking behaviour as inquired in Weitkunat et al. (2013). Likewise, each participant consented to be video recorded during the study procedure traceability purposes.

[13] Access to the *MAT* experimenter app: `https://github.com/MiezelKat/LabMAT`
[14] As meditation music we used song number 14 from the album '72 Ambient Meditations'

Figure 5.7.: Experimental Setup in the Body-Centric Lab of Queen Mary University London. Participants were either seated or walked on a treadmill in front of the screen presenting the *MAT* exercises.

After, participants were given a short introduction to the treadmill and the *MAT* was explained. Following, the participants were asked to put on the chest-worn *ECG* sensors of the *Nexus* and the *Polar H7* chest belt. A sheet with visual material from the manufacturers and instructions on the correct sensor placement were provided to ensure proper sensor fit (Appendix E, Figure A.10a). Those illustrations where discussed with the participants, before they proceeded. Participants were instructed to place the *Polar H7* chest belt around the chest and close the clasp after moistening the plastic parts at the inside of the strap. The figure also depicts the colour-encoded placement of the pre-gelled *ECG* electrodes. Participants were also instructed to already put the cables for the skin temperature and skin conductance through the sleeve of their right arm, to prepare for the following steps. The participants were given privacy behind a screen to perform these actions. The correct placement of the sensors and the correctness of the *ECG* signal was later checked on the experimenter's computer.

After the chest sensors were placed, the experimenter placed the skin conductance sensors on the fingers and attached the skin temperature sensor on the participant's forearm using medical tape. Next, participants were aided in placing the wrist-worn devices (*Apple Watch* and *Microsoft Band*). Before the study started, correct data transmission and validity of the signals for all sensors was checked. Each procedure starting with the recording of the baseline. In this primary condition all participants were asked to remain seated for five minutes listening to meditation music via wireless headphones. Then, each participant was assigned the following conditions in counterbalanced order, alternating between *walking* and *stationary* while *MATs* should be performed. Between the two stress-inducing trials, there was a fixed task requiring to walk while listening to meditation music via wireless headphones.

To assess the *dependent variables (DVs)*, *SAM* questionnaire including single-items on wake/tense arousal and perceived stressfulness of the task were administered after each trial (including the baseline). The entire procedure took approximately 1.5 hours in total. The aforementioned discussed material is included in Appendix Appendix E.

5.2.4. Data Collection and Analysis

To collect sensing data from the consumer devices, the LabExperiment app (Section 4.2.3, Appendix B) has been used. The data from the *Nexus* was obtained using the *BioTrace+* software by MindMedia (2017). A detailed overview of the collected sampling rates of the data was discussed in Section 4.2.1. During an experiment session and after the sensors have been put on by the participant, the experimenter checked the correct data transmission by using the *BioTrace+* Software and the LabExperiment app. The LabExperiment app had feedback on the heart rate readings of all three consumer test devices - namely *Apple Watch*, *Microsoft Band*, and *Polar H7*. It was checked that these readings are in a similar range and change over time. The fit of the *Nexus* sensors, and especially the *ECG* electrodes, was checked with manual inspection within the *BioTrace+* software. Special emphasis was put on the shape of the *ECG* signal and the "spikes" characterising heart beats[15]. To synchronise across the Nexus and the consumer devices, manual markers were placed at the same time within the *BioTrace+* and LabExperiment app. At the same time, the experimenter vocally indicated the marker placement for the video

[15] A visualisation of a typical *ECG* signal can be found in Section 2.2.2.1

recording, to synchronise the video feed. Throughout the experiment, markers have been placed when an experiment condition began and ended.

5.2.4.1. Preprocessing

The data obtained from the *BioTrace+* software and the LabExperiment app were exported in CSV files and stored in folders per participant. Participant IDs were hereby anonymised. Sensor data and questionnaire data was held strictly separate from any identifiable information of the participants and was stored on an encrypted external hard drive.

Missing Data. For 3 participants (P14, P18, P20), there was a problem with the signal from the *Polar H7* device. For the purpose of power savings, the hear rate chest strap is fitted with a sensor detecting if the device is worn. As soon as the device is loosing contact with the skin, data transmission is stopped and the device is turned off. It is turned back on, when skin contact is reestablished. Unfortunately, once the data transmission was stopped, the data recording for the *Polar H7* was discontinued within the LabExperiment app. It was not reestablished even when the *Polar H7* was transmitting data again. For P14 and P18 this happened early on in the experiment and there is no data available for any of the four experiment conditions. For P20, just the first experiment condition was completed. Accordingly, the *Polar H7* data was excluded for these 3 participants.

Furthermore, there was a bug in the LabExperiment app for the *Polar H7* data recording. While the profile for the *RR-Interval* readings could contain a list of readings, always solely the first one was read. This lead to the potential loss of readings and therefore *RR-Interval* and consequently *Heart Rate Variability (HRV)* analysis was excluded for this experiment.

Separation of the Data into Condition Windows. For the analysis, the data was split into separate files per condition. A period of 4 minutes per condition was taken. The first 40 and last 10 seconds were omitted due to novelty effects.

Adjustment of Measurement Units. The *Nexus* and *Microsoft Band* use different approaches to represent *EDA*. The *Microsoft Band* uses skin resistance, while the *Nexus* provides skin conductance. To convert the Microsoft Band's provided skin resistance measures (resistance R in $kohms$) to match the unit provided by the Nexus device (conductivity G in $micro - mho$),

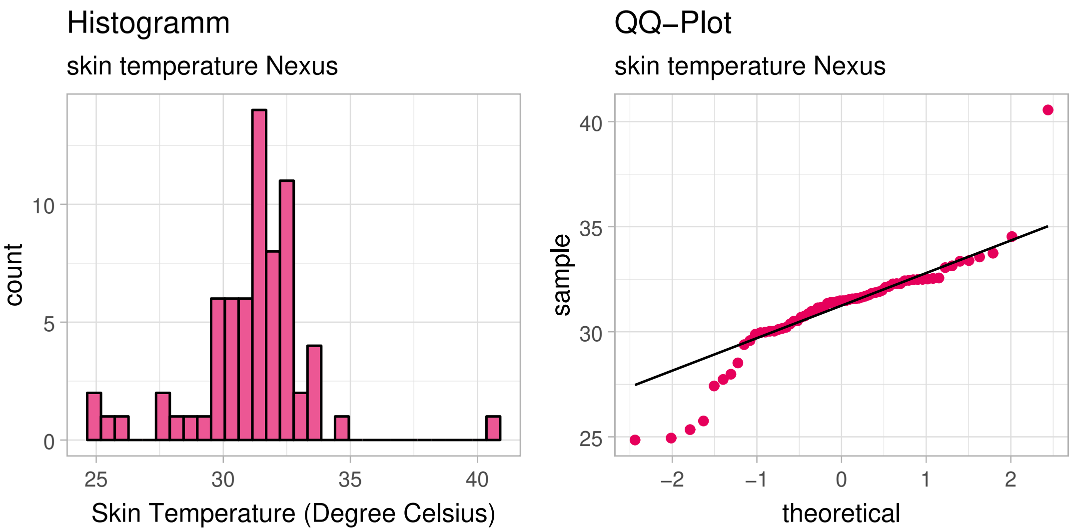


Figure 5.8.: Histogram and QQ-Plot to exemplarily show the non-normal distribution of skin temperature sensing data of the *Nexus* (Shapiro-Wilk test rejected the null-hypothesis of normal distributed data ($W = 0.868, p - value = .000$))

the following formula was applied:

$$G = \frac{1}{R} * 1000$$

5.2.5. Analysis

To investigate the Hypotheses from Section 5.2.2, inferential statistics was used. The analysis was performed using SPSS and R.

Normal-Distribution of the Data. To decide on the appropriate test method, the normal-distribution of the data was investigated. The normal distribution was tested using *Shapiro–Wilk test* and the visual inspection of Histograms and QQ-Plots. A sample of histogram and QQ-Plot and histogram for the *Nexus*'s skin temperature data can be found in Figure 5.8 whereby the Shapiro-Wilk test rejected the null-hypothesis of normal distributed data ($W = 0.868, p - value = .000$).

Due to the determined non-normal distribution of the data, non-parametric statistical tools were used.

Error Rate. As per Hypothesis 5-5c, the error rate of all sensor signals compared to the reference device (*Nexus*) was calculated. This was done for every device d and sensor source s within a 5 second time window w. For every window w the mean of sensing data readings ($\bar{r}$) were taken and compared to the *Nexus* by calculating the percentage of difference:

$$error_{d,s,w} = \frac{|\bar{r}_{\text{Nexus},s,w} - \bar{r}_{d,s,w}|}{\bar{r}_{\text{Nexus},s,w}} * 100$$

Overview of the Data. An overview of the data in form of boxplots is presented in Figure 5.10. The descriptive measures (Mean, Median, Standard Deviation) for the recorded physiological data by each device and the subjective measures grouped per condition (relaxed stationary (RS), MAT stationary (MS), relaxed walking (RW), and MAT walking (MW)) is presented in Appendix Appendix G.

5.2.6. The Impact of Stress

This first part of the results focuses on the impact of stress on the subjective and sensor measures. To summarise, expected outcomes were:

- an increase in subjective stress measures (arousal, tense arousal, wake arousal, perceived stress) under the *MAT* conditions (Hypothesis 5-3).
- changes in physiological measures - namely increase in heart rate, increase in skin conductance, and decrease in skin temperature - measured by the different devices (Hypotheses 5-5b and 5-4b).

5.2.6.1. Stress and Subjective Measures

As proof of concept, that the stimulus material in form of relaxation music and *MAT* induced low and high stress, the subjective stress ratings were considered. This was in accordance with Hypothesis 5-3 that subjectively perceived arousal and stress are higher after a stress task was performed. The results indicate that subjective stress was induced in the *MAT* conditions. *Wilcoxon-Signed-Rank test* with a Holm-Bonferroni sequential correction were performed. Table 5.3 shows the results.

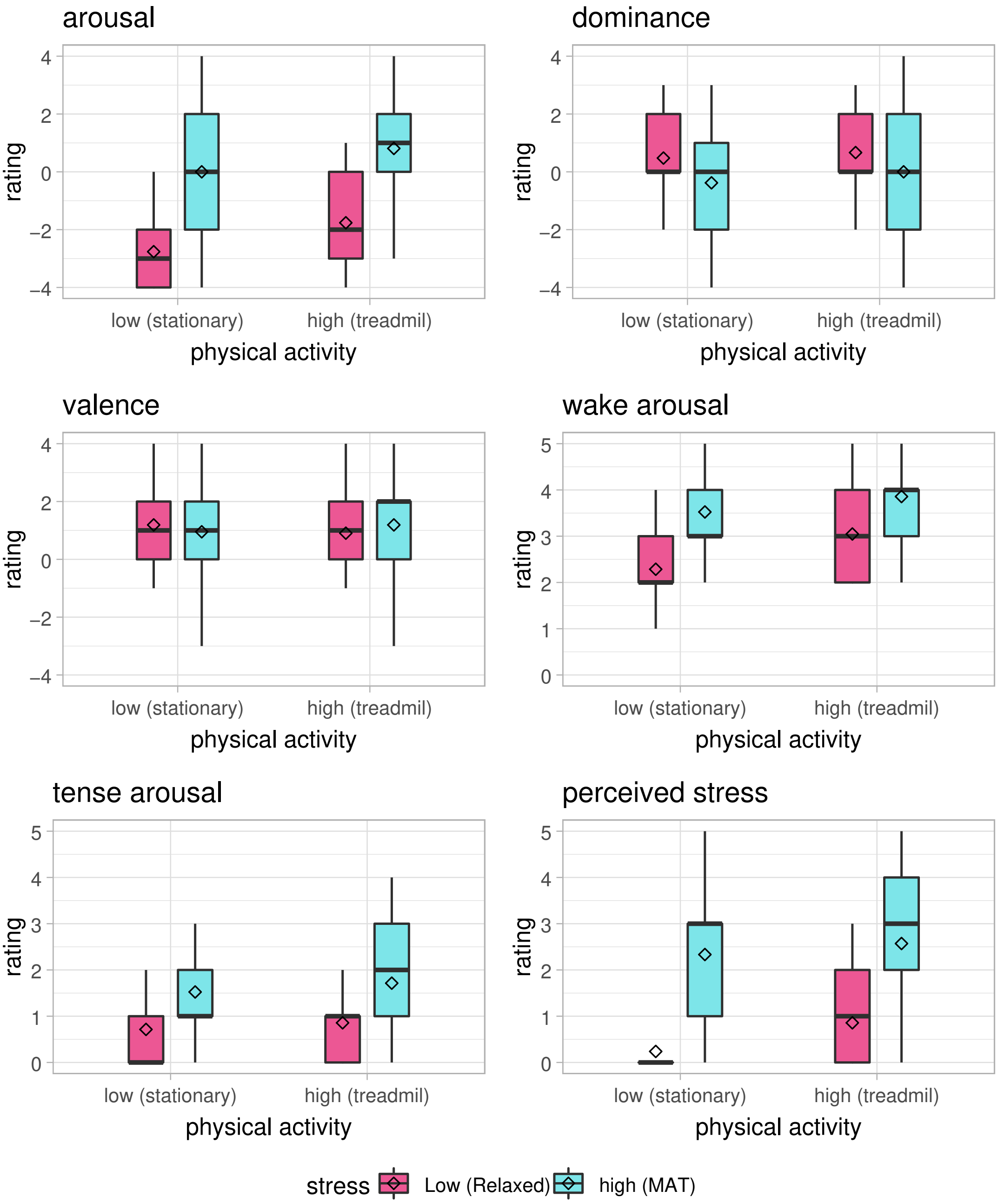


Figure 5.9.: Boxplots (including mean (⋄)) of the subjective measures grouped per levels of physical activity (x-axis) and stress (color).

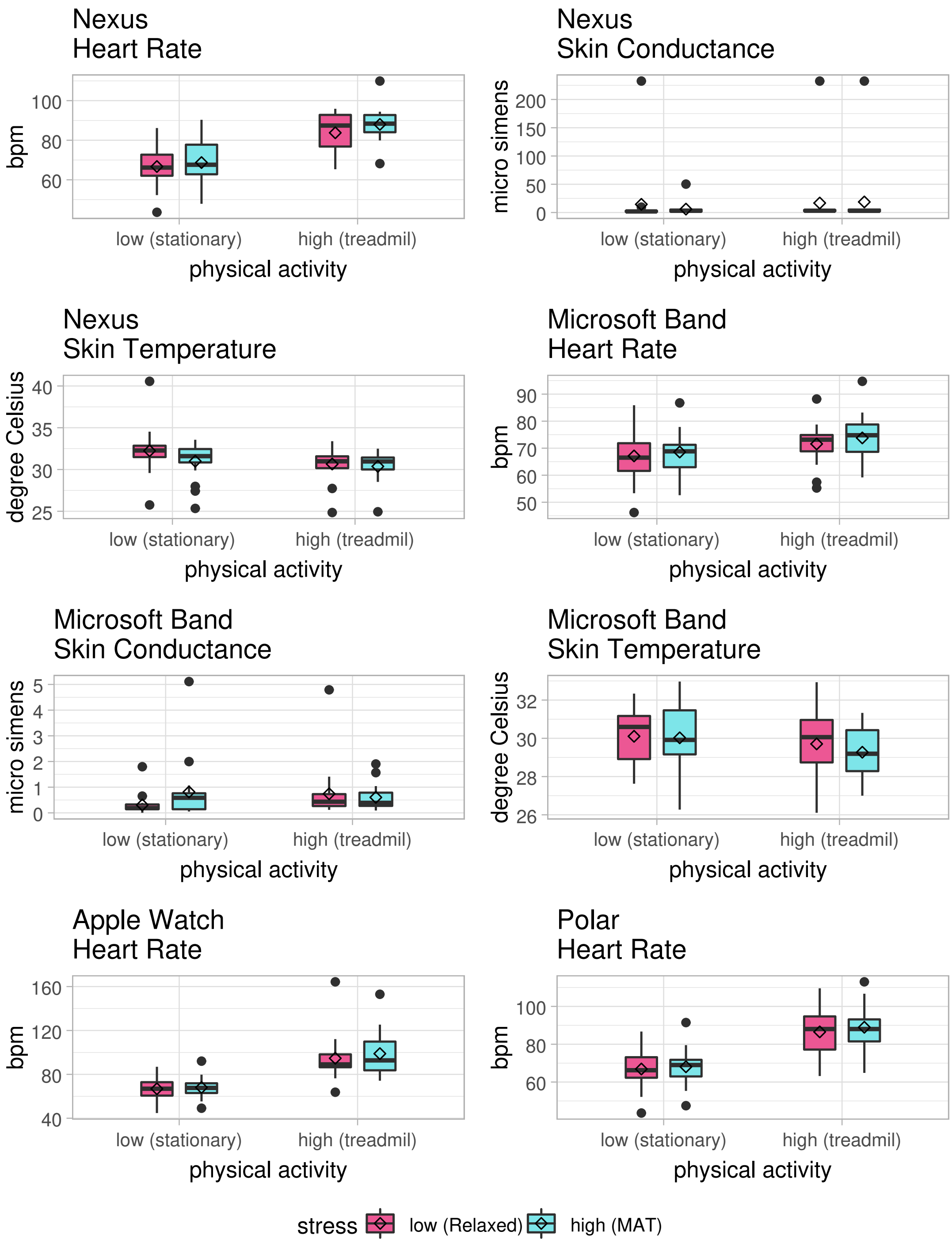


Figure 5.10.: Boxplots (including mean (⋄)) of the sensor measures grouped per levels of physical activity (x-axis) and stress (color).

There was a significant higher arousal, tense arousal, wake arousal, and perceived stress during the stationary (Table 5.3a) and treadmil (Table 5.3b) *MAT* conditions. There were no differences observed for dominance and valence ratings, indicating that they are not significantly affected by the stress inducing task. This confirms the Hypothesis made.

5.2.6.2. Stress and Physiological Measures

The core part of the experiment focused on the devices ability to pick up the hypothesised changes in physiological markers; namely an increase in heart rate, skin conductance, and decrease in skin temperature under stress. The cases of low physical activity and high physical activity were considered separately in Hypotheses 5-4a and 5-4a. Changes were investigated using Bonferroni corrected *Wilcoxon-Signed-Rank test* for the non-normal distributed data (see section 5.2.5. These test were performed on each device's sensor source. For the two planned comparisons the Bonferroni correction on $\alpha = 0.05$ which resulted in $\alpha/2 = .025$.

Those changes in sensor signals were confirmed under low physical activity for the *Nexus*; it showed a significant increase in heart rate ($Z = -2.381, p = .017$), increase in skin conductance ($Z = -3.285, p = .001$), and decrease in skin temperature ($Z = -2.416, p = 0.016$). The *Microsoft Band* showed a significant decrease in skin conductance ($Z = -3.058, p = 0.002$). The other sensor sources and devices did not show any significant changes depending on the stress levels under low physical activity.

Regarding the high physical activity conditions on the treadmill: There were no significant changes to report for all devices and sensor sources.

5.2.7. The Impact of Physical Activity

Following the Hypotheses 5-5a and 5-5b that movement and subsequently physical activity has an impact on the accuracy of physiological sensing readings, the *IV* of *physical activity* has been included in the experiment setting. But physical activity in itself can be seen as a physical stressor on the body and has an impact on the physiological responses of the body, e.g., an increase in heart rate under physical activity (see Hypotheses 5-6a and 5-6b). Following, the

	Relaxed		MAT		Wilcoxon-S. R. Test	
	Mean (SD)	Median	Mean (SD)	Median	Z	p
Arousal	-2.76 (1.14)	-3	0 (0)	2.28	-3.580	.000†
Tense A.	0.71 (1.01)	0	1.52 (1.17)	1	-2194	.028†
Wake A.	2.29 (1.06)	2	3.53 (0.93)	3	-3.095	.002†
Dominance	0.48 (1.6)	0	-0.38 (2.31)	0	-1.422	.155
Valence	1.19 (1.47)	1	0.95 (1.69)	1	-.820	.412
Perceived Stress	0.24 (0.54)	0	2.33 (1.39)	3	-3.095	.002†

(a) Stationary conditions

	Relaxed		MAT		Wilcoxon-S. R. Test	
	Mean (SD)	Median	Mean (SD)	Median	Z	Significance
Arousal	-1.76 (1.64)	-2	0.81 (1)	1.99	-3.647	.000†
Tense A.	0.86 (0.73)	1	1.71 (1.19)	2	-2.994	.003†
Wake A.	3.05 (0.97)	3	3.86 (0.79)	4	-3.494	.000†
Dominance	0.67 (1.32)	0	0.0 (2.32)	0	-1.186	.236
Valence	0.9 (1.14)	1	1.19 (1.63)	2	-.948	.343
Perceived Stress	0.86 (0.96)	1	2.57 (1.4)	3	-3.688	.000†

(b) Treadmill conditions

Table 5.3.: Descriptive statistics and results of the post-hoc pairwise Wilcoxon Signed Rank Test for the subjective measures amongst relaxed and MAT conditions. Figure a) shows the comparison during the stationary activity and Figure b) during the treadmill activity. The significance level with Holm's sequential Bonferroni is marked with †.

effects of physical activity on sensing data reliability and physiological/subjective measures are discussed.

5.2.7.1. Activity and Sensing Readings Across Devices

It was hypothesised that sensor readings are fairly consistent across devices under stationary activity (Hypothesis 5-5b), but under high physical activity show increased incongruence across devices (Hypothesis 5-5a) and higher error rates compared to the reference device (Hypothesis 5-5c).

Heart Rate. Firstly, the heart rate across the four devices were investigated. A Friedman test revealed no differences across the devices under low physical activity when participants were sitting still; $\chi^2 = 4.286, p = .232$. This goes in accordance with the hypothesis, that devices are fairly accurate and consistent amongst each other under low physical activity and movement.

In contrast, there were significant differences under high physical activity; $\chi^2 = 43.133, p = .000$. Post-hoc tests showed a difference between the pairings with the *Microsoft Band*. The *Microsoft Band* (MSB) reported a significant lower heart rate compared to the *Nexus* (N), *Polar H7* (P) and *Apple Watch* (AW); $Z_{N,MSB} = -4.773, p = .000$; $Z_{P,MSB} = -4.583, p = .000$; $Z_{AW,MSB} = -4.156, p = .000$. This highlights that the *Microsoft Band* shows higher incongruences with the other devices.

Skin Temperature. Skin temperature readings were provided by both the *Nexus* and *Microsoft Band*. The *Wilcoxon-Signed-Rank test* indicated a significant difference between the Nexus and Microsoft Band among both physical activity conditions alike. The Microsoft Band showed a lower skin temperature in general regarding stationary condition ($Z = -4.503, p = .000$) and walking condition ($Z = -4.256, p = .000$). On average, the *Microsoft Band*'s reported skin temperature value was 1.31°C (Mdn = 1.40°C; $\sigma = 2.32$°C) lower than the *Nexus*' skin temperature over all conditions. This difference and variation can be explained with the sensor placement at different body positions.

Skin Conductance. Skin conductance readings were provided by both the *Nexus* and *Microsoft Band*. The *Wilcoxon-Signed-Rank test* revealed a significant difference of *EDA* measures between the *Nexus* and *Microsoft Band* among both physical activity conditions. The *Microsoft Band* showed a lower skin conductance in general with respect to the stationary condition - $Z = -5.125, p = .000$

		Heart Rate			Skin Temperature	Skin Conductance
		Polar	Apple Watch	Microsoft Band	Microsoft Band	Microsoft Band
overall	Mean	6.84	8.28	12.06	6.63	9751.29
	SD	12.34	15.52	12.04	6.77	39889.37
stationary	Mean	3.22	3.42	5.44	7.32	10295.95
	SD	4.07	4.12	5.96	5.73	43214.71
walking	Mean	10.28	14.41	19.03	5.79	9178.71
	SD	16.03	21.37	12.87	5.15	36079.91
Mann-Whitney	U	48	39	3	253	213
U Test	p	.000	.000	.000	.271	.949

Table 5.4.: Average error percentage of the sensor signals compared to the *Nexus* reference device. The error rates are once shown over all conditions and separated into low physical activity (stationary) and high physical activity (walking).

and walking condition - $Z = -5.024, p = .000$. Here again, the *Microsoft Band*'s reported *EDA* was 11.817 Micro-Mho (Mdn = 2.500 Micro-Mho; σ = 44.188 Micro-Mho) lower on average than the *Nexus*' *EDA* over all conditions. This difference in skin conductance can be explained with the different sensor placement and distance between electrodes[16].

Error Rate. Comparing the error rates, to the laboratory measurement instrument - *Nexus* - revealed further differences among the two different physical activity conditions. The mean, standard deviation, and statistical tests of differences between stationary and walking conditions are presented in Table 5.4.

Considering the average heart error rates, the *Polar H7* chest belt performed best with an average error of 6.84%. The *Microsoft Band*'s average error was 12.06% over all conditions. Considering

[16]The conductivity of a material is indirect proportional to distance between the measuring electrodes; a shorter distance means a higher conductivity.

the stationary and walking conditions separately, all devices showed a significant higher error in heart rate readings during the walking conditions[17].

Skin temperature readings of the *Nexus* and *Microsoft Band* did not differ significantly under stationary (mean error 7.32%) compared to walking (mean error 5.79%) conditions; $U = 253, p = .271$. This highlights that while there is an error, it was not significantly affected by movement.

Similarly, the error rate of the skin conductance readings of the *Microsoft Band* did not differ significantly under stationary and walking conditions; $U = 213, p = .949$. The overall error rate is very high (9751.29%) due to the high difference in readings between the devices. As mentioned before, this high difference in sensor values can be explained by the different placement and distance between electrodes.

In favour of our hypothesis, the error was higher in the walking conditions for all the heart rate sensors. The *Microsoft Band*'s skin conductance and skin temperature sensors showed a consistent error rate in stationary and walking conditions.

5.2.7.2. Activity and Physiological/ Subjective Measures

Physical activity causes changes within our body, such as increases in heart rate, sweating, etc. These changes were considered using *Wilcoxon-Signed-Rank test* with a Holm-Bonferroni correction. The results of this test are presented within Table 5.5a.

For the relaxed conditions, there was an observed significant increase in heart rate, skin temperature and *EDA* in all devices apart from the *Microsoft Band*'s *EDA* measure. Similar changes were expected for the *MAT* conditions; but the *Microsoft Band*'s *EDA*, heart rate and skin temperature measures and the *Nexus*' *EDA* and skin temperature sensors did not show these changes.

Similar tests were performed for the subjective measures (see Table 5.5b). Neither in the relaxed conditions nor in the *MAT* conditions these measures showed a statistical difference between the two levels of physical activity apart from a difference in tense arousal during the relaxed conditions; there tense arousal was significantly higher during the treadmill condition (mean

[17] The difference in error readings across the data sources during stationary and walking condxitions were determined with a *Mann–Whitney U test* test

		Heart Rate				**Skin Temperature**		**EDA**	
		Nexus	Polar	Micros. Band	Apple Watch	Nexus	Micros. Band	Nexus	Micros. Band
	Z	-3.806	-3.724	-2.240	-4.015	-4.015	-1.717	-3.980	-3.920
R	p	.000	.000	.024	.000	.000	NS	.000	.000
	$p^{\dagger}$	.000	.000	.048	.000	.000	NS	.000	.000
	Z	-4.015	-3.724	-2.213	-3.920	-2.450	-1.730	-1.547	-0.161
M	p	.000	.000	.027	.000	.014	NS	NS	NS
	$p^{\dagger}$	.000	.000	NS	.000	NS	NS	NS	NS

(a) sensor measures

		arousal	valence	dominance	tense arousal	wake arousal	perceived stress
	Z	-2.417	1.223	-0.652	-0.629	-2.669	-2.446
R	p	.014	NS	NS	NS	.005	.014
	$p^{\dagger}$	NS	NS	NS	NS	.030	NS
	Z	-1.632	-0.764	-0.445	-0.533	-1.538	-0.767
M	p	NS	NS	NS	NS	NS	NS
	$p^{\dagger}$	NS	NS	NS	NS	NS	NS

(b) subjective measures

Table 5.5.: These tables presents results of the *Wilcoxon-Signed-Rank test* between the two physical activity levels (stationary and walking) once for the relaxed (R) and *MAT* (M) levels. Significant results of the *Wilcoxon-Signed-Rank test* indicate a difference in physiological signals measured by the device. Changes where hypothesised for all signals and devices. (NS - Not Significant, $\dagger$ - Holm-Bonferroni adjusted p-values)

tense arousal of 3.047) compared to the stationary conditions (mean tense arousal of 2.285) indicating that participants felt more awake when being physically active. This indicates that while physiological measures changed the subjective measures showed no differenced depending on the physical activity level.

5.2.8. The Relationship between Subjective and Physiological Measures

As hypothesised in Hypotheses 5-7a to 5-7d, there is an expected relationship between subjective measures – namely arousal, valence, and perceived stress, but not dominance – and physiological sensor measures. The significance threshold for the Spearman correlations on the non-normally distributed data was set to $p < 0.05$. To control for individual differences in the participants' heart rate, skin temperature and *EDA* responses, the physiological data was transformed using within-subject z-score standardisation, as suggested by Ben-Shakhar (1985).

The heart rate provided by *Nexus*, *Apple Watch* and *Polar H7* showed correlations with perceived arousal, wake arousal and stress. The *Microsoft Band*'s heart rate showed mere correlation with the arousal measure. On the contrary, while both *EDA* sensor data showed a weak agreement with arousal, wake arousal and stress, the *Microsoft Band*'s *EDA* measure additionally, and as the only sensor source, correlated with tense arousal. The reference device's skin temperature measure correlated negatively with the arousal, wake arousal and tense arousal, while the *Microsoft Band* showed mere correlations of skin temperature with wake arousal. Contrary to the hypothesis 5-7c, none of the physiological data sources showed correlations with valence. Additionally, hypothesis 5-7d that dominance does not correlate with physiological data was confirmed. The results of the correlation are presented in Table 5.6.

5.2.9. Summary

This section summarises the previously presented results, limitation, and take-aways of this laboratory experiment.

Validity of the Stress-inducing Task. As a stress-inducing task, *MAT* have been chosen. This task set was shown to induce subjective and physiological stress responses (Callister et al., 1992).

		arousal	**wake arousal**	**tense arousal**	**perceiv. stress**
Heart Rate	AW	.265*	.244*	-	.252*
	MSB	.244*	-	-	-
	Polar	.235*	.236*	-	.248*
	Nexus	.323**	.277*	-	.284**
EDA	MSB	.361**	.272*	.337**	.376**
	Nexus	.297**	.362**	-	.296**
Skin Temp	MSB	-	-.221*	-	-
	Nexus	-.259*	-.367**	-	-.262*

Table 5.6.: Correlation of subjective measures and within-subject normalised physiological data from Nexus, Polar, Microsoft Band (MSB) and Apple Watch (AW); (** Correlation is significant at the level $p < 0.01$, * Correlation is significant at the level $p < 0.05$, omitted values (-) were non-significant)

Within this study, the perceived arousal and stress levels were increased during the *MAT* and an increase in heart rate, *Galvanic Skin Response (GSR)*, and a decrease in skin temperature was observed; this matches the expected and hypothesised responses. This confirms the proof of concept of the study setup.

5.2.9.1. Wearables as Predictors of Stress

The statistical tests on differences within the stress-inducing *MAT* and the relaxing tasks was considered. The analysis showed that merely the professional device *Nexus* was able to detect changes in heart rate, skin temperature, and *EDA* during the stress task. This effect was just observed during the stationary conditions, when participants were seated. None of the consumer devices were able to pick up those changes, apart from the *Microsoft Band*'s *EDA* sensor during the stationary conditions.

During the walking conditions, there were no statistically significant changes observed. Additionally, the significantly increased error rate heart rate sensors 5.2.7.1 indicates that the movement during physical activity disrupts sensor readings. This highlights the difficulty to use physiological signals to predict stress, since there is uncertainty on the reliability of the sensor sources.

Furthermore, changes within physiological signals are similar for cognitive stress, as it is induced by the *MAT*, and physiological stress induced by the walking task; both of these induce an increase in heart rate, increase in skin conductance and decrease in skin temperature. This makes it harder to predict the source and cause of the physiological changes. This makes it harder to identify the underlying cause of physiological changes. Additional sensing measures can help giving context of the current situation; e.g., accelerometers can be used to detect movements to not just get context on the user's current activity, but also to make assumptions on the sensor data's accuracy.

5.2.9.2. Validity of Sensing Devices

As shown by related work and within this laboratory study, movement and physical activity has an impact on the sensors' validity. The heart rate error percentage of all the test devices compared to the *Nexus* increased significantly under physical activity. The *Microsoft Band*'s error percentage was highest with almost 20% while participants were walking.

Oppositely, the *EDA* and skin temperature sensors, showed an error but it was fairly consistent; no significant changes were reported between the stationary and physical activity conditions. The reported error can hereby be explained with the sensor placement. The skin temperature sensor of the *Microsoft Band* was placed at the wrist[18] protected from external temperatures through the wearable itself. The *Nexus*' temperature sensor was placed with adhesive tape on the inner, upper forearm of the participants. This made it more susceptible to external temperatures[19]. Moreover, skin surface temperature can differ for different body segments. This can explain the tendency of the *Nexus* to report a higher skin temperature with an average of 1.31°C (±2.32°C).

[18]The exact location of the *Microsoft Band*'s skin temperature sensor is not disclosed by the manufacturer.
[19]The temperature within the Body-Centric lab was recorded for each participant session and was 21.79°C (±0.68°C) over the course of the 10-day study.

Similarly, *EDA* and skin conductivity/resistance are highly dependent on the placement of the electrodes and especially the distance between the electrodes. In physics, the resistance changes proportionally with the length of the conducting material. *EDA* sensors measure the resistance and conductivity of the skin surface, hence a larger distance between the measuring electrodes increases the length of the way the current has to travel. Summarising, a higher distance means a higher resistance and lower conductivity[20].

This indicates, that the skin temperature and *EDA* sensors of the *Microsoft Band* and *Nexus* are more robust and less prone to movement, since the error rate is consistent during low physical and high physical activity.

5.2.9.3. Discussion

The effect shown in this study is not novel: consumer devices have shown to be less reliable opposed to professional devices in terms of heart rate accuracy[21] (El-Amrawy and Nounou, 2015; Pietilä et al., 2017). Especially motion and movement have been shown to increase error and approaches to counteract this trend have been introduced. Especially filtering mechanisms to remove motion artefacts have been explored, e.g., moving average filters or spectral filters (Tamura et al., 2014; Wood and Asada, 2007). These are approaches applied to the *PPG* signal itself and therefore, they are not applicable to most consumer devices available on the market, since they commonly do not offer access to this signal; it can just be speculated that device manufacturers themselves already apply filtering approaches.

For stress recognition outside of the lab, hybrid approaches of contextualising the wearable physiological data with other behavioural or environmental data have been applied. For example, Sano and Picard (2013) used binary classification of stress with wearable (accelerometer and *EDA*) and mobile usage data with an accuracy of 75%. Mozos et al. (2017) used wearable skin conductance and *PPG* (BioPac sensor) and sociometric badge to classify stress with 67% of accuracy. Gjoreski et al. (2017) presented a sophisticated approach of using context in form of accelerometer-derived physical activities (e.g., standing, walking) and wearable physiological data from the *Empatica E4 wristband*; adding this contextual information resulted in an improved

[20] Conductivity σ is the inverse of resistance R: $\sigma = \frac{1}{R}$
[21] To the thesis author's knowledge, there is no work evaluating other physiological sensors in consumer goods, i.e., skin temperature/conductance.

stress recognition from a mean F-score of 0.47 (without context) to a mean F-score of 0.9 (with activity context).

5.2.9.4. Limitations

Although our results of this study are giving important insights into the reliability of physiological data accessed by wearables, only a limited amount of devices were tested. Facing the variety of wearable (fitness) devices, the results may not apply for each of them. The presented devices cover a range of different heart rate sensing technologies, such as *ECG* in the *Polar H7* chest strap, stick-on, 4-lead electrode *ECG* in the *Nexus* and two *PPG* wrist sensing technologies; still, the impact of various sensing hardware and software mechanisms for data collection and processing mean that the results are not generalisable. Further, the reliability of wrist-worn *PPG* heart rate sensors is influenced by factors, like skin pigmentation (Spierer et al., 2015), which have not been assessed during the study. Additionally, the results are based on a short-term data acquisition of approximately 20 minutes; further long-term validation of the device performance in a longitudinal setting also including more participants could provide deeper insights.

Since all participants were students with engineering background, there are implications on the performance during the *MAT*. Although the subjectively assessed measures indicate that participants felt more stressed in the *MAT* conditions, task performance was not analysed. A further investigation of participants' task performance and the adaptive adjustment of the *MAT*'s difficulty would be interesting to observe also with respect to subjective and physiological stress measures. The unfortunate loss of *RR-Interval* data from the *Polar H7* device, made the *HRV* analysis unfeasible. This analysis could have given deeper insights in stress reactions of participants.

5.3. Discussion

Within this chapter, two research probes have been presented. Firstly, the *Apple Watch* sensing data has been used to show a relationship with affect experience samples collected in the wild. Secondly, four wearable devices — one professional grade and three consumer devices — have been evaluated for their suitability to detect stress and affect changes. In the following, the results and implications of these two research studies are discussed.

5.3.1. Reliability of the Data

Especially the laboratory study showed a difference between the four exemplary wearable devices. While the *Nexus 10 MK2*, which was chosen as reference due to it being advertised as a professionally sensing device, was the only device to show the hypothesised stress changes. Further, the error rates of the devices varied greatly and increased when participants where physically active.

5.3.2. Charging of the Devices

As has been discussed before (Section 4.4), collecting and transmitting sensing data has implications on the battery life of a device. None of the participants using the *EmoRate* app in their daily lives, reported a negative impact on the battery life. However, the experimenter experienced a quite remarkable difference in how often the devices needed to be charged during the laboratory experiment (Section 5.2). One of the advertised benefits of most wrist-worn devices is their mobility aspect. Without the need of cables, they allow the unconstrained movement of the participant. Additionally, their relatively long battery lifetime allows for them to be worn for a long time without the need to charge. While the Apple Watch promises a an 'all-day' battery life of 18 hours and the Microsoft Band 48 hours (Apple Inc., 2017; Andronico, 2017), the Polar provides 400 hours of heart rate recording (Polar Support, 2017).

The Nexus promises more than 24 hours of operation (Mind Media, 2017). All of the devices are advertised as wearable, but the Nexus would hardly be suitable for e.g. sleep studies, due to its bulky nature.

5.3.3. Participant's Comfort

Although, the comfort of putting the devices on the participants has not been formally assessed, there were issues reported by participants. Some participants, when removing the disposable, sticky electrodes of the *Nexus*, reported discomfort from the glue. Additionally, when putting on the *Electrocardiogram (ECG)* chest sensors of the *Nexus* device, participants were presented with a handout on where to place the sensors so they could be put on the sensors by themselves behind a privacy screen. Still, some participants asked the experimenter for help to place the electrodes. This could potentially lead to uncomfortable situations and this should be kept in mind by researchers.

On the other hand, the wrist devices *Apple Watch* and *Apple Watch* were naturally easy to attach and it is to assume that they provide higher comfort when worn than the bulky *Nexus* or *Polar H7* device. After all, these consumer smart and fitness watches are designed to be worn all day and even night.

5.3.4. Implications for using Wearable Devices in Research

Within the two presented research probes, four devices have been used. These are the same devices that have been assessed for their means to obtain data in the previous Chapter 4. Important factors such as the granularity/nature of the accessible data and technological means to obtain data have been discussed. This discussion focuses on further aspects uncovered during the execution of the studies. The following paragraphs discuss important aspects and questions researchers should ask themselves when using a device in a research setup.

Validity and Reliability of the Data. What has been highlighted within the laboratory study is a discrepancy in the device's suitabilities for sensing stress. Factors such as movement and activity have been shown to increase the unreliability and error rate of sensing data. Researchers have to be aware of these when choosing a device for a research setup.

The questions researchers need to consider are:

a) What level of accuracy does the phenomenon to be measured require? - e.g., *Autonomous Nervous System (ANS)* responses

b) Are factors present during the study which further impact a devices reliability? - e.g., physical activity and movement has been widely shown to affect device reliability

c) What research studies have shown the reliability and suitability of the chosen device in similar setups?

Impact of Wearing a Device on Participants. The mere attachment and wearing of a device can have an impact on the participants. This has to be considered when choosing a device.

The questions researchers need to consider are:

a) Is the device suitable and safe to be worn in general? — e.g., self-developed devices should be safe for participants to wear and should have been tested rigorously before exposing study participants; in case of a commercial device, is it certified for the intended use

b) How comfortable is it to wear the device for the rough length of the proposed study setup? — e.g., participants may not feel comfortable to wear a restrictive device for several days

c) Are there restrictions in participants to attach the devices? — e.g., most common wrist devices may not be suitable for children's smaller wrist sizes, or vulnerable participants may not be comfortable with intrusive sensors being attached

b) Could wearing the device cause a bias in the collected data itself? — e.g.,

Device Features. In some cases, devices have specific features which are relevant, these need to be considered. For example:

a) Is there special feedback from the device itself needed? — e.g., haptic, auditory or visual feedback

b) Are special input modalities required? — e.g., programmable *user interface (UI)* with touch screen or buttons

c) Should the device be operational under certain circumstances? — e.g., during swimming

Remote Experiments. Furthermore, the experimenter - i.e. the thesis author - experienced several issues in the process of preparation, execution and analysis of the experiment.

Part IV.

Conceptualisation

Chapter 6

Design Space for Wearable Measurement Tools

This chapter ties together the findings from the previous two exploratory chapters to conceptualise a Design Space for Physiological Measurement tools. The original design space has been previously published by Hänsel et al. *(2018b), and the evaluated form is currently under review (*Pogunktke et al.*, 2018).*

The previous two chapters explored how suitable data can be obtained from wearable devices (Chapter 4) and how this data can be used for affect sensing in the lab and in the wild (Chapter 5). The findings were discussed in terms of their implications for researchers. Within the overall theme of this thesis to discuss the suitability of wearable sensing devices for their application in research settings, it has been established that there is no one-size-fits all answer. Rather, there are several criteria which can be applied to specific devices to make assumptions on their suitability for certain settings. This chapter summaries and generalises these findings to answer RQ 3: *What are criteria for choosing a suitable wearable sensing devices in a research settings?*

To answer this question, this chapter presents a design space for choosing wearable sensing devices in research. Firstly, an *Initial Design Space (IDS)* is presented. It is derived from the reflection of the last two chapters and the experiences with the four test devices. This design space contains the following five dimensions[1]:

Data richness describes the multitude of sensors available and the granularity of the data (e.g., sampling rates, precision)

[1]The design space in this initial form has been published a contribution by Hänsel et al. (2018b)

Data accessibility describes the the effort and means needed to gain access to the data for the processing and analysis

Data reliability considers the accuracy and validity of the data for the given purpose.

Mobility considers factors such as battery life, restriction through cables which make a device less suitable for a mobile use

Comfort of attachment considers how comfortable a device is to attach, wear, and detach

Since these dimensions are based on the reflection on the experience with the four test devices and the two conducted research probes, more evidence is needed to validate this *Initial Design Space*. Consequently, this chapter presents this evidence in form of a qualitative study where 10 interviews — 5 interviews with experts using wearables in their research, and 5 consumers using wearables in their daily lives — were conducted. The results confirm the five dimensions presented above, allowed the derivation of 2 more dimensions and 19 sub-dimensions.

6.1. Initial Design Space

As explored in Chapters 4 and 5, there are several factors impacting the suitability of wearables to be applied in research. This *Initial Design Space (IDS)* is an approach to present dimensions which are crucial when evaluating and choosing a wearable device for a research setup.

In Chapter 4, it has been uncovered that there are several different architectures on how data access is provided by device manufacturers and that granularity and level of preprocessing differs greatly. As exemplary shown with the four test devices, the professional device *Nexus 10 MK2* offered easy access to sensing data with high sampling rates. On the contrary, the consumer devices did not offer straight-forward data access; a special app needed to be developed to obtain the data from these devices[2].

[2]LabExperiment app is discussed in Section 4.2.3

In Chapter 5, two research probes were presented. Conducting these two studies uncovered further aspects where the four test devices differed, e.g., the reliability of the data. The following sections describe the *IDS* dimensions and argue the placement of the four experiment devices within this space.

6.1.1. Data Richness

As described in Section 4.2, the four test devices chosen are devices with data sources relevant for affect sensing. While the set of sensors are comparable, the granularity and sampling rates of the provided data differs greatly. Considering the heart rate measures, the *Nexus* provided a raw *Electrocardiogram (ECG)* signal with a frequency of 256 Hz, while the *Polar H7 ECG* chest belt did not allow access to the raw signal but merely the heart rate and *RR-Intervals*. On the contrary, the *Apple Watch* provided roughly one heart rate sample every couple of seconds and the *Microsoft Band 2*'s heart rate sampling rate is 1 Hz.

Not just the granularity of a device is important, but also the diversity of sensors. The *Microsoft Band* is particularly richly equipped for a consumer device with e.g. heart rate, skin temperature, *Electrodermal Activity (EDA)*, and UV sensors compared to other wrist-worn wearables.

6.1.2. Data Accessibility

Not just the availability and richness of a sensor reading, but also the effort to access the data is important. In Section 4.2.2, the means of collecting and exporting the data from the four test devices were discussed. Differences exist from where, i.e., the device itself, the mobile gateway, or online cloud services, and with how much effort data can be accessed.

The BioTrace+ software suite, which accompanies the *Nexus*, provides easy export and even real-time data visualisations making access easy and allows low-effort checks of the data during an experiment. Apple included *HealthKit* in their iOS system which allows *Comma Separated Values (CSV)* export of the collected heart rate samples. The *Polar H7* and *Microsoft Band* sensor data is mere accessible through mobile *Application Programming Interfaces (APIs)*, which have to be included in a data collection app, or third-party applications. Here it becomes obvious that the ease of data accessibility needs to be improved.

6.1.3. Comfort of Attachment

Comfort or wearability of wearables are not just an important factors for acceptance of the device (Bodine and Gemperle, 2003), but play an important role for the study device choice; a device that is clumsy, uncomfortable and restricts participants can have an impact on the naturalness of behaviour, influence the mood or affective state, and, depending on the research question, corrupt the validity of the data for the given purpose; e.g., if a device frustrates participants then this has an impact on task performance and sensor readings and can induce potential bias (Hokanson and Burgess, 1964).

The wrist wearables are designed to be worn all day long and have a high social acceptance due to their placement on a natural location (Profita et al., 2013). Hence, they are suitable for long-term in-situ studies, like the *EmoRate* study presented in Section 5.1. The *Nexus* and *Polar H7* are more purpose-led in their functionality and are designed to be worn for certain occasions, like running or biofeedback session. The chest belt form factor of the *Polar H7* device makes it suitable for, e.g., field studies due to its easy and quick attachment. But it can be visible through tight-fit clothing and may not be comfortable, especially for female participants, due to its placement. The *Nexus*, as a laboratory measurement tool with several applications, is relatively heavy (500 grams[3]) and requires detailed instructions on the correct placement of sensors. Therefore, it is cumbersome for research settings requiring flexibility and unobtrusiveness. Further, the self-adhesive stick on electrodes can cause discomfort when removed and may leave behind residue.

[3]approximated weight by the manufacturer: `www.mindmedia.info/CMS2014/products/systems/nexus-10-mkii` (accessed: 02/01/2018)

6.1.4. Mobility

Mobility can play an important role when restrictions through the device can impact the study data, compliance, or feasibility. Such restrictions can be, e.g., cables which restrict movement, a short battery life, or a limited transmission range (in case there is no internal data recording mechanism). Mobility is especially important in long-term and in-the-wild experiments, where participants do not want to be restricted in their behaviour over a longer time. But even during shorter laboratory experiments it has to be taken into account how a device restricts participants or researchers, e.g., how often devices need to be charged needs to be considered for the experiment schedule.

A huge benefit of most wrist-worn devices is their mobility aspect. Without the need of cables, they allow the unconstrained movement of the participant. Additionally, their relatively long battery lifetime allows for them to be worn for a long time without the need to charge. While the Apple Watch promises a an 'all-day' battery life of 18 hours and the Microsoft Band 48 hours (Apple Inc., 2017; Andronico, 2017), the Polar provides 400 hours of heart rate recording (Polar Support, 2017). The Nexus promises more than 24 hours of operation (Mind Media, 2017). All of the devices are advertised as wearable, but the Nexus would hardly be suitable for e.g. sleep studies, due to its bulky nature.

6.1.5. Data Reliability

Most of all, the previously presented studies and related work (Section 3.2) confirmed that there are variations in sensor data accuracy which results in a limited reliability.

The *Nexus* device, as a laboratory tool, was the only device to show stress-related, statistically valid changes in heart rate, skin temperature and *EDA* in the laboratory experiment. Wrist-worn, *Photoplethysmography (PPG)*-based devices tend to be less reliable in measuring heart rate than devices deriving heart rate values from *ECG* data. This effect is worsened in conditions involving physical movement. But there are even differences amongst devices using *PPG* technology. The *Microsoft Band* was identified to be the most unreliable in terms of heart rate and skin temperature data while the *Apple Watch* performed acceptable. On the contrary, while heart rate

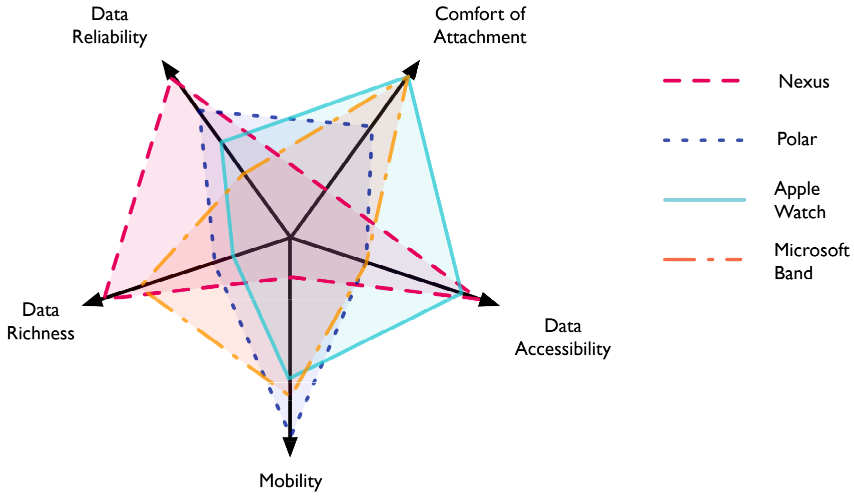


Figure 6.1.: Illustrative schematic of the design space evaluation for our 4 test devices (Nexus, Polar, Apple Watch, Microsoft Band) in 5 criteria dimensions (data reliability, comfort of attachment, mobility, data richness, and data accessibility)

chest belts with *ECG* technology proved to be more reliable than *PPG* sensors (i.e. (Gillinov et al., 2017)), there were no significant differences in sensing data between stressed and relaxed conditions with the Polar device in the laboratory study.

6.1.6. Summarising the advantages and Fallbacks of the Test Devices

Considering the four named dimensions of the discussed devices, the fulfilment of each criteria per each device is illustratively depicted in Figure 6.1.

Nexus 10 MK2. As shown within the laboratory study, the *Nexus* device offers very fine-grained, high-frequency, easily accessible and reliable data; its clear benefits are the *richness*, *reliability* and *accessibility* of the data. With a high sampling rate and as the only device able to detect subtle changes during stress inducing tasks, it is well suited for studies on monitoring *Autonomous Nervous System (ANS)* responses and affective states of study participants. But its bulkiness and the cables required to connect the sensors to the base device, impact the *mobility* and the *comfort* for the participants. Furthermore, it can cause discomfort by requiring the researcher to place sensors at intimate body parts, such as the chest. It is a laboratory device which is best suited for settings which require not much movement and physical activity from participants. It can serve as a reliable measurement tool for stationary studies in the lab.

Wrist-worn wearables. The *Apple Watch* and *Microsoft Band* are wrist devices designed to be worn all day long. Their form factor makes them well-suited for non-stationary settings and field studies. They offer a high mobility with a battery life of at least a day and no restrictive cables; the suitability for long-term studies can be improved by choosing a device with longer battery lives. Further, these devices were designed for a long-term comfort for consumers and their high *comfort of attachment* makes them suitable for studies with vulnerable subjects where more-intrusive devices may be unaccaptable, e.g., children, elderly or disabled people. The drawbacks of both the *Apple Watch* and *Microsoft Band* lie in their accuracy; as this laboratory study showed, both devices are less suitable for studies where subtle reactions in physiological signals have to be detected. On top of this, there are differences in how the data can be accessed from both devices and how fine-grained the data is. Due to the large variation in terms of the accessibility, reliability and granularity of sensing data in this device category, the decision for using a wrist-worn device has to be made on a case by case basis. It is recommended, that the chosen device is evaluated against a reference device, like, e.g., the *Nexus*, before its application in a study. In general, these wrist-devices would be suitable for longitudinal, in-the-wild study which focus on physiological changes with a higher magnitude, e.g., physical activity related metrics.

Polar H7. The *Polar H7* chest strap forms a viable, and also less expensive alternative to the other devices. The differing technology on sensing heart rate with *ECG* leads to a lower error rate compared to the *Microsoft Band* and *Apple Watch*. The drawback lie in *data richness* since it only assesses heart rate and *RR-Interval*; depending on the study setup this may be enough. Additionally, the data from the device is just accessible though third-party apps or by integrating the bluetooth protocol into a proprietary app to collect the data (see 4.2). The *Polar H7* is the least expensive device used in this study. Its form factor makes it less suitable for long-term studies but its high mobility would make it suitable for field studies.

Concluding, researchers should weigh the pros and cons for utilising the discussed sensing technologies considering study setup, flexibility needed and purpose of the study. Thorough testing and evaluation in smaller pilot studies is recommended before a device is deployed in larger studies. In the following sections, a qualitative evaluation of these dimensions is presented.

6.2. Qualitative Evaluation of the Initial Design Space

The subjectivity of the previously discussed *Initial Design Space (IDS)* requires the need for proper validation to consequently and thoroughly answer the research question of this chapter. A qualitative study to gain deeper insights on important criteria for choosing wearable measurement devices in research is proposed. This study investigated the experiences and opinions of established and experienced researchers — following named *experts* — using semi-structured interviews. To avoid bias from the *IDS*, the focus is on firstly collecting unbiased opinions on experiences and choice criteria for wearable sensing data in research settings. Secondly and within the framework of the previously developed *IDS*, an opinion and ranking of the proposed five dimensions and reasoning behind the ranking was investigated. To summarise, the aim of the evaluative qualitative study is to a) explore unbiased criteria on how researchers choose wearable confirm and evaluate the 5 *IDS* dimensions (Section 6.1), and b) explore novel dimensions and aspects of choosing a wearable measurement device in research setups.

6.2.1. Method and Procedure

To evaluate the design space in its initial form, five interviews with researchers were conducted[4]. These researchers are work in the field of affective computing and are with experience of using physiological and wearable sensing. Semi-structured interviews as proposed by Bernard (2006) were conducted. These interviews followed an interview guide (presented in Section 6.2.3) with proposed questions and topics. The average interview lengths for *experts* (24.3 min, SD = 2.52 min) and *consumers* (24.17 min, DS = 6.53 min) were comparable. The interviews

[4] The qualitative study additionally included interviews with five wearable device owners — consumers. These will not be reported here, since they do not contribute to the research questions of this chapter.

were audio-recorded and transcribed. The ATLAS.ti[5] software was used for the annotation of the interview transcripts. The annotation was followed by a thematic analysis adopting a process of developing from lower level 'codes' in a first annotation round towards higher level 'themes' (Braun and Clarke, 2006). The ethical approval letter and consent form can be found in Appendix H.

6.2.2. Interviewee profile

The interviewees were convenience-sampled through personal contacts or academic online profiles. The five expert interviewees (M = 30.6 years, SD = 2.61 years) were selected based on a) the years they work in their field of research, b) the variety of wearable sensing devices they worked with, and c) the number of peer-reviewed publications involving physiology sensing. This resulted in a sample of researchers who worked with at least 3 different sensing devices for at least 4 years with several publications using sensing from wearables (M = 11.2, SD = 5.2). Details on their demographics and experience can be found in the Appendix I (Table A.2).

6.2.3. Interview structure

The semi-structured interviews followed the approach of Bernard (2006); an interview guide with topics and proposed questions was provided to the interviewer to explore how researchers use wearable devices in their research. The main aim was to validate the proposed 5 *IDS* dimensions for choosing a wearable sensing device in research settings. The guide was divided into three sections: a) unbiased opinion and experiences, b) opinion on the *IDS* dimensions, and c) new perspectives and closing remarks from the interviewees.

Unbiased Opinion. To gather unbiased and genuine opinions on crucial criteria and experiences of research experts when using wearable measurement devices, no information about the *IDS* was given during the first part of the interviews. The interviewees were asked to freely talk about their experience when using devices in their research and what criteria they have for choosing a device.

[5] `www.atlasti.com`, Accessed: 26/04/2019

Evaluation of the Initial Design Space (IDS). In the second part, the interviewees were presented with an overview of the five dimensions of the *IDS* (see participant material in Appendix H, Figure A.12). The dimensions were presented in a circle to avoid any bias introduced through, e.g, an ordered list. Interviewees were first enquired to clarify the understanding of the five dimensions and what they mean for them. Following, the interviewees were asked to provide a ranking and reasoning for the ranking.

New perspectives and closing. Finally, all interviewees were asked if they have any other comments or opinions not covered before. It was also enquired if they have any other opinions.

6.2.4. Interview analysis

The analysis of the transcribed interviews was conducted using ATLAS.ti software following the six-step approach by Braun and Clarke (2006). This approach comprises: a) familiarisation with the data, b) generation of initial codes, c) searching for potential themes, d) review of these themes, e) defining and naming final themes, and f) writing of the report. Two independent coders, followed this approach and generated the initial codes and potential themes. This followed the inductive approach for the first part of the interviews on unbiased opinions, and the last part on new perspectives and closing remarks. The second part of the interviews was based on evaluating the *IDS* dimensions and therefore a deductive top-down analysis was performed. The five dimensions acted as codes for the interview analysis, and new, not fitting aspects were labelled with new open codes. The merged codes were then compared and mismatches between the two coders were extensively discussed to reach a consensus. From these codes, new themes were derived from the codes and reviewed by the two involved researchers. The themes were also discussed against the concepts and dimensions of the *IDS*; this led to two new dimensions in the design space and 19 sub-dimensions.

6.3. Unbiased Experiences with Wearable Sensing Devices

The first part of the interviews focused on the unbiased opinion on experiences and decision criteria for using wearables in research. The aim was to find important criteria for choosing a suitable measurement device for a research setup without introducing bias through the *Initial Design Space (IDS)*. A inductive bottom-up approach was followed whereby novel concepts are extracted from the data without any preexisting framework or hypothesis. These concepts were then summarised to higher level concepts or dimensions. Below, the emerging observations are discussed.

Importance of Sensing Data. It was apparent that the properties and aspects of the sensing data of a device are naturally crucial for researchers. All of the five experts mentioned aspects on the quality of the data, such as '*noise*' (E1, E5), data being '*unreliable*' (E2), '*measurement quality*' (E3), '*data validity*' (E4), or '*contradictions with another device*' (E5). It became apparent from the interviews that quality of the data is an important aspect of wearable devices, but it is also problematic. Multiple interviewees have expressed their feeling that there are shortcomings with one or more devices they used. The element of '*trial and error*' (E2) seems to be prevalent. Especially, E5 mentioned that sometimes he '*had the feeling that the values are not sufficiently accurate or inaccurate*' (E5) and E1 '*believes*' that fitness trackers have noise in their data. This highlights the problem that sometimes researchers cannot be sure about the reliability and accuracy of a devices data; thus, they sometimes rely on very subjective 'feelings' to make a judgement. According to E3, devices become more '*trustworthy*' the more papers have been published with it. Which highlights that the feeling of trust in a device can be improved through more publications or, e.g., through '*clinical studies that confirm[] that it [, the device, has] almost a clinical accuracy*' (E5).

The term 'accuracy' has different meanings for different interviewees. E4 defines accuracy as the '*precision of the current measured value with respect to the actual value. The signal that you get from a clinical sensing tool can have an accuracy of 1000 Hz but the commercial sensors can be a 1Hz signal*' (E4); this can be more interpreted in terms of the data resolution or the precision of the data. On the contrary, E5 considers lack of accuracy as '*data that do not represent the reality and that you cannot rely on.*'

The '*amount*' (E1) and '*relevance*' (E2, E3) of the collected sensing data as well as a sufficient '*sampling rate*' (E3, E4) have also been highlighted. The problematic that for '*many of them [wearable devices] the API doesn't give you access to the raw data all the time*' (E2) has been acknowledged.

Consumer Vs Clinical Sensors. E4 highlighted the bad experience he had with consumer sensors; in his opinion '*most of the consumer version sensors, are either averaging a lot [...] and that is not as accurate as you get from clinical sensors*' (E4) Other experts (E1,E2) shared this experiences, when using devices such as fitness trackers or smartwatches, describing them as '*not fulfilling*' (E2), '*very unreliable [data]*' (E2), or with '*noise in their data*' (E1). On the other hand, E1 works mainly on affective computing algorithms in the wild and he '*want[s] to use devices that are not a burden to use in everyday life*' (E1); consumer goods could address this issue. Additionally, many consumer devices provide companion apps which allow the visualisation of the data. This was one criteria for E3 to chose the Fitbit because it provides a '*nice analysis that the participants can see. So, they benefit from using this sensor*' (E3). In her opinion, '*it [the Fitbit] is also more comfortable*' (E3). The positive experience with the manufacturer support, when using the professional *Empatica E4 wristband* was mentioned by E5. This is another factor to consider when using consumer devices which are not sold for the research purpose; the technical support could not be helpful in these cases.

Operation of a Device. The factor of a device needing to be '*straight-forward*' (E4) and '*easy to synchronise*' (E3). E4 explains that he wants to '*know when it has started, and that it [the device] would indicate clearly it has actually started*' (E4). E5 described the difficulty with non-documented devices, which make the use harder: '*the [serial] interface was badly documented or not documented at all*' (E5). These reports highlight, that usability issues and poorly documented tools can negatively impact their usage.

Focus on the User. While from this initial set of questions, the aspects mentioned covered mainly the data, some interviewees mentioned user related criteria. In particular, E3, who is working in the field of *Human-Computer Interaction (HCI)* and interruptibility, mentioned as the very first thing that it is important whether '*it [the device] would be restrictive for the participants in their daily life or if it is easy to wear*' (E3). Similarly, E2 who stated that when working on *HCI* applications, he '*needed to not set up people with bulky hardware, even if it is high resolution, [...] because [he] was looking into concepts that work in the real world for a normal user*' (E2). Both statements explicitly highlight 'daily life' and 'real world' studies where the properties around restrictiveness and bulkiness of a device are naturally important.

Choice of a Device. When asked on how the researchers choose a device for their research, the most common thing mentioned at first was 'data', and '*for scientific use, the most important feature [] is getting a very accurate value, with a high time resolution*' (E4). The interviewees also expressed that the choice is not easy and that there are '*trade-offs*' (E3) for the devices. The choice of a devices often depends on the '*study context*' (E3) and '*type of study*' (E5). This summarises that there is no ultimate answer on what devices are appropriate in general, but rather the needs of the study need to be considered.

Summary. The above mentioned aspects were mentioned by the interviewees before the design space was introduced. They were also analysed with an inductive approach, where these annotations and codes where generated without the influence of the *IDS*. Still, in the final discussion it became apparent that the emerging aspects of this unbiased part of the interview were assignable to the Design Space Dimensions. Aspects matching *IDS* dimensions have been mentioned by at least two experts (Table 6.1). It is not just apparent that the dimensions are valid, but also that the initial focus of researchers lay on the devices provided data and the quality. All experts mentioned that the reliability (concerning the data quality but also the reliability of the device) was crucial when choosing a device to work with in research studies. This was more so the case for laboratory studies where reliability was valued as more important than comfort of a device; for field studies the opposite was true. Aspects of considering the comfort to attach a device to a participant and its mobility aspects in general were less often mentioned. The aspects of how trustworthy devices manufacturers are or factors such as the device operation were also less prevalently mentioned.

In general, this highlights that there naturally is no one-size-fits-all solution. All the various aspects and themes mentioned by the researchers have been summarised below to build the Design Space for Physiological Measurement Tools.

		E1	E2	E3	E4	E5	count
Initial Design Space Dimensions	Comfort of Attachment		•	•			2
	Data Accessibility		•	•	•	•	4
	Reliability	•	•	•	•	•	5
	Data Richness	•	•	•	•		4
	Mobility		•	•			2
New Dimensions	Operability				•		1
	Trustworthiness			•		•	2

Table 6.1.: Overview of the expert interviewees mentioning (•) of top-level design space dimensions during the first part of the interview (before they have been presented with the dimensions of the *IDS*)

6.4. Evaluation and Extension of the Design Space

In the second part of the interviews, participants were presented with the five dimensions of the *Initial Design Space (IDS)*. It was enquired what these dimensions refer to and how important these are for participants. Apart from the five already mentioned dimensions of the *IDS*, two additional dimensions arose: *Trustworthiness* and *Operability*. Additionally, each interviewees was asked to rank the five *IDS* dimensions. These ranks are summarised in Table 6.2

6.4.1. Comfort of Attachment

The placement of wearable sensors can have an impact on participants. The dimension of *Comfort of Attachment* addresses these issues. As E3 summarises, this comprises that '*[a] sensor is comfortable to wear, unobtrusive no matter where you wear it, [and] it should not impede.*' (F)rom this statement, the following sub-dimensions can be derived: *Comfort of Wearing*, *Unobtrusiveness*, and *Degree of Restrictiveness*. The *Comfort of Attatchment* is greatly influenced by the *Physical Properties*, such as '*bulky hardware*' (E2). To go even further, devices should not be invasive and even cause harm, e.g., '*skin irritations*' (E3); this leads to the dimension *Invasiveness*. *Comfort*

	E1	E2	E3	E4	E5
Comfort of Attachment	2*	1*	1*	5	3
Mobility	2*	1*	-	3	4
Data Richness	4	5	1*	4	5
Data Reliability	1	3	1*	1	1
Data Accessibility	5	4	-	2	2

Table 6.2.: Ranking of the *IDS* dimensions. The higher the rank, the higher the priority. *indicates when dimensions were explicitly ranked as equally important by the interviewees.

of Attachment was also linked to the participants behaviour: *'The sensor needs to not affect the behaviour of the participant, and that they wouldn't feel they are putting something on.'* (E4) This highlights that invasiveness of a sensor is not just important for ethical reasons — as mentioned by E5 — but also the implications on the user's behaviour and respectively the collected data.

6.4.2. Mobility

The dimension of *mobility* was often associated with restrictions, e.g., from wires: *'For the mobility, we wouldn't want the users to be stretched to a wire, but I guess we would go with something that is wireless and allows mobility'* (E1). Further a devices ability to connect to other devices, e.g., the phone (E2). This led to the derivation of the new sub-dimensions: *Wirelesseness*. E4 remarked that *Mobility* is closely related to the *Comfort of Attachment* from a users perspective and that sensors should be *'unobtrusive when attached to the body'* (E4). Both E4 and E5 mentioned the *Ease of Setup*: *'how easy are these things [devices] to transport, how easy is the setup, how easy is the carrying'* (E5). *Mobility* received a mixed ranking. While some interviewees ranked it as very important (E1, E2), others considered it less so (E4).

6.4.3. Data Richness

The dimension *Data Richness* initially described the number of sensors and the granularity of sensing data. E1 associates *Data Richness* with the variety in sensors in a device and argued: *'we* *always prefer the devices that have the most amount of sensors in the least amount of space'* (.) Similar statements were made by E2 (*'multiple types of data'* ()) and E4 (*'how many different signals are recorded'* ()) being summed up in the sub-dimension *Sensor Variety*. Apart from the variety of the sensors, the characteristics of the data are also important. Several interviewees mentioned the significance of *'high resolution'* (E2), *'precision or time resolution'* (E4), and *'sampling rate'* (E3). This resulted in the sub-dimension *Degree of Resolution*. As discussed in Section 4.1, many wearables do not just offer primary sensing data, but also processed features, e.g., step counts derived from motion sensor data. The dimension *Degree of Preprocessing* reffers to the availability of *'not only one measurement parameter but more derived measurement parameters'* (E5).

Still, when asked to rank the five IDS dimensions, *Data Richness* was not considered as too important. Apart from E3, who stated *'reliability, comfort of attachment and data richness are forming a combination that can be considered as "minimum criteria"'* (E3)

6.4.4. Data Accessibility

With respect to the *Data Accessibility* E1 clearly stated that he would *'prefer devices which give open access to the data [...] [and] also you don't have to develop your own algorithm for Bluetooth transmission of the data'* (.) This highlights that *Data Accessibility* mainly revolves around the effort needed to obtain said data. Multiple interviewees (E1, E3, E5) stated that the need to *'code something'* (w)ithout the data access *'work[ing] out-of-the-box'* (E3) is an issue. Additionally, the lack of documentation can make this task more difficult (E5). These aspects have been summarised in the sub-dimension *Data Accessibility Effort*. Related to this aspect is the *Data Transmission Effort* incorporates the temporal component whether data can be transmitted and accessed when needed. This included real-time data transmission and *Connectivity*, as mentioned by E1 and E5, or the need for manual synchronisation to transfer the data to the desired destination: *'How do I access the data - Do I have to be on site and transfer the data to something or can they [the participants] do it by themselves or does it happen automatically?'* (E3). E5 highlighted the importance of the *Data Format* and that data *'should not be a proprietary format by the company that one cannot read out by oneself or interpret [but] it should be a format that is commonly accepted e.g. a CSV file'* (.)

6.4.5. Reliability

Referring to the IDS dimension *Data Reliability*, E1 said: '*this is the most important subject that you have here on this list because if you are developing machine learning algorithms which would work with the data given by sensors, then the machine learning algorithms would learn also some noise in the sensory data*' (.) Likewise E3 and E5 highlighted that '*the data we get should measure what we want and do that accurately*' (E5) and '*sufficiently accurate regarding measurement quality*' (E3) referring to *Data Quality*. When asked to rank the five IDS dimensions, *Data Reliability* was considered the most important one by the majority.

E3 and E4 further mention the reliability regarding the recording and software in the new sub-dimension *Software Reliability* by phrasing the criteria '*how sure I can be that the data is not corrupt, and the data measurement starts at the right moment, and that it stop at the moment when I want it to stop*' (E4). Since these two dimensions are both characterising how reliable a device's sensing hardware and software operate, they have been summarised as sub-dimensions of *Reliability*.

6.4.6. New Dimension: Trustworthiness

Also the availability of the devices pointed out by E5 was one of the criteria when choosing a device. Because he considered that, e.g., '*often are these things [are] announced for a long time but then they are not published [...] or it is not possible to ship it to Europe*' (E5). This aspect has not been considered in the IDS and, hence, it is added as the additional dimension *Trustworthiness*. The sub-dimension *Availability* characterises how easy it is to obtain a device from a manufacturer and how long it is likely to be supported to avoid experiences such as E3's: '*support was stopped, server were down, so sensors became useless*' (.)

Likewise, the sub-dimension *Degree of Testedness* describes the confidence in the companies or other researchers that a device has been thoroughly tested. E3 asks herself: '*"what publications do exist with this device?" so that it becomes trustworthy. [...] [She] looked if it is a standardised, validated sensor or something experimental*' (.) E5 also mentioned that the existence of clinical studies can improve the trust in a device.

6.4.7. New Dimension: Operability

Another new dimension not considered before is the *Operability*. This dimension on one hand comprises the *Clarity* when operating a device, e.g., '*It is important for me that when I start using the tool for sensing, I know when it has started, and that it would indicate clearly it has actually started*' (E4). On the other hand, the *Ease* describes that a device is easy to operate, e.g., '*easy to start and stop and only limited interaction is needed*' (E3).

6.4.8. Additional Themes

Additionally to the above mentioned dimensions, interviewees also mentioned facets applying to wearables in research which are included in the seven design space dimensions due to their subjectivity.

Privacy. E2 mentioned the aspect of privacy which is crucial especially for users. With many wearable architectures, especially in the consumer sector, transferring data to proprietary cloud services (see Chapter 4), using such devices for studies implies that data is not just visible to the research team.

Social Acceptability. E2 mentioned the aspect of a device being socially acceptable and '*that user are fine with wearing it in public*' (E3). While this aspect is related to the dimension *Comfort of Attachment*, it has not been included as a dimension or sub-dimension due to its subjectivity. While E2 goes into details on what is acceptable, i.e., '*wrist bands are acceptable; rings that look odd are acceptable because they are small; eye glasses are starting to be a bit accessible, but they are a little bit freaky in public*' (,) this opinion is very subjective to a) the study setup and b) the perception of the wearer and audience, and c) social setups.

Financial Cost. The cost of obtaining a device was mentioned by 4 interviewees (E1, E2, E3, E5). It has been stated that there is a tradeoff between a reliable device and cost, e.g., '*we have one project, and it is really important to have reliable data and the device that can be easily bought, so we got the budget to buy 50 Empaticas, and it was not a problem*' (E1); In a different project with less finances, they bought Microsoft Bands due to the cheaper cost. E3 indicates that sometimes the affordability of a device is an added convenience. She admits that it is a '*question whether we*

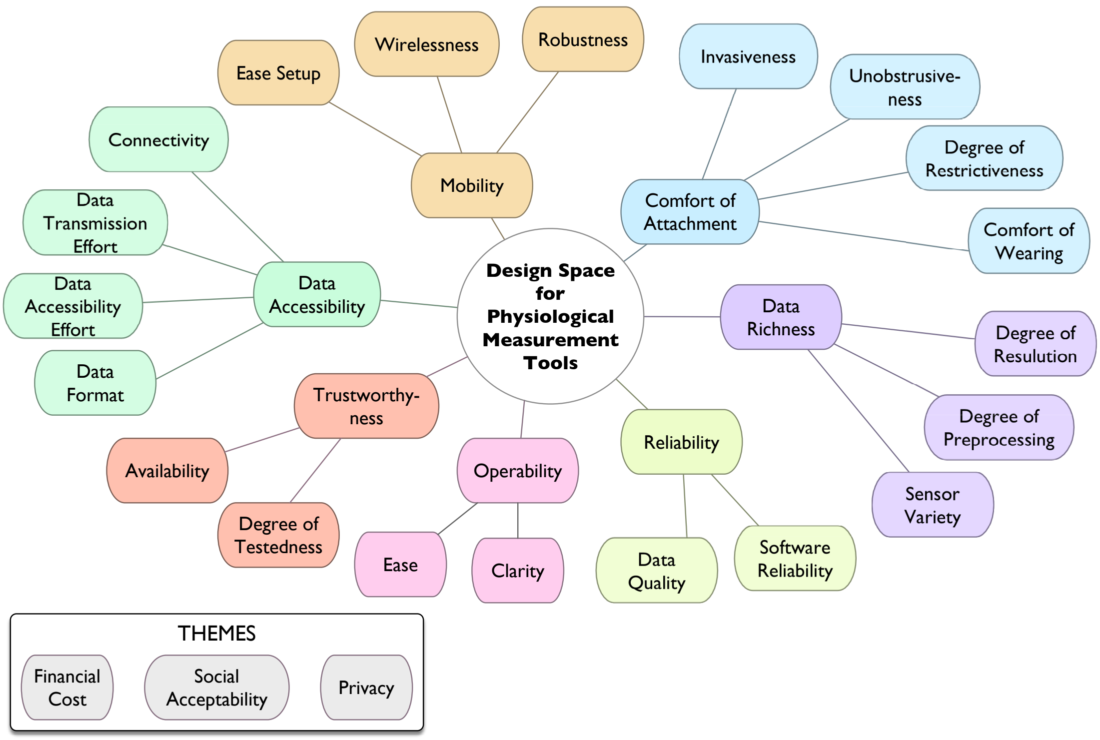


Figure 6.2.: Design Space Dimensions

want to have more devices and use them parallel or have a limited number of devices and use them maybe sequentially, [with the] consequence that it takes longer' (E3). According to E2, not just the device but also costs for software licenses have to be considered. While this is a popular and crucial aspect. The financial affordability of a device for research is highly dependent on subjective factors, such as research funding, the quantity of devices required, the availability of a device within the department, or the option to borrow the device.

6.4.9. Summary

With the above mentioned dimensions, the design space can be extended to 7 dimensions and 20 sub-dimensions. Further, there are several aspects which are not distinct dimensions but rather very study and context dependent things to consider. An overview of the design space is depicted in Figure 6.2.

6.5. Discussion

This study presented an *Initial Design Space (IDS)* on criteria for choosing measurement devices in research. This *IDS* was evaluated with semi-structured interviews to form a more complete design space with 7 dimensions and 20 sub-dimensions.

6.5.1. Design Recommendations

During the expert interviews, several recommendations for the design and execution of research studies with wearables have been mentioned.

Participant Inclusion. Keeping participants engaged in research can be crucial; Especially, when studies are longer. While many participants chose altruistic reasons for participating in research, other benefits, such as insights, are also present (Hayman et al., 2001). Providing participants with actual 'value' in form of insights can provide an additional incentive in keeping them engaged. Users value to monitor their own data and activities (Packer et al., 2014). E3 made the observation, that it can be beneficial to provide '*a benefit for the participants if they wear the device. For example for the Fitbit, they see the analysis and so they are more interested to participate; They find it exciting*' (E3). Wearables and their data have been an extensive subject of research regarding the behaviour change their data can provide to users (Hänsel et al., 2015; Mercer et al., 2016). When designing user studies, researchers should consider to provide valuable insights and information to participants; provided that doing so does not pose a bias or confounding variable.

Increase Trust. As shown in 6.3, many participants expressed the feeling that devices are not very reliable in terms of the data they provide. Important is the conveyed uncertainty in these expressions. E5 stated that he relied on '*clinical studies that confirmed that it [device] had almost a clinical accuracy*' (E5). In conjunction with recommendations by Evenson et al. (2015) and Shcherbina et al. (2017), standardised evaluations can help to make a more informed statement on the data reliability of a device.

Handling and Usability. An aspect mentioned by multiple interviewees was the reliability and feedback of the device itself. According to *E2*, '*sometimes you do a long session, and then you realise* *a lot of data was corrupt, so it [the device] sensed incorrectly*' (E2) Similarly, E5 reported: *"I do not want to realise in the afternoon that it stopped recording after half an hour"*E5. During the laboratory experiment in Section 5.2, the *Polar H7* chest strap lost skin contact for several participants and consequently turned itself off. This went unnoticed for several cases, because there was no feedback integrated in the device. In research, loosing participant data due to technical difficulties is frustrating. The 'Visibility of system status' is one of the 10 usability heuristics for user interface design proposed by Nielsen (1994). This design recommendation is rather universal and can mainly be addressed by device manufacturers.

6.5.2. Application of the Design Space

The presented design space presents 7 dimensions of criteria for choosing a wearable device for research settings. When choosing a wearable device, researchers should a) prioritise these, and b) evaluate wearables accordingly. The application of the design space to a novel device has been presented in the next chapter, Section 7.2.2.

Prioritisation of the Dimensions. The priority and importance of each dimension can not universally be determined. As shown in Section 6.4, Table 6.2, the importance has been ranked differently by the experts. E4 explains that in her opinion, there is a set of minimum criteria: '*On the one hand, [the device] needs to be sufficiently comfortable so that participants are willing to wear it. And there is an individual threshold, when it is okay or not okay. Regarding the data, if the data is crap then the device can be as comfortable as possible but it does not lead anywhere. So the basic prerequisite needs to be given for all of them: so that one can access the data somehow, that [the device] provides you with what you need, with the accuracy that I need for my processing, and sufficiently comfortable, and not restrictive that people are willing to participate*' (E4).

On the researchers side, data reliability, which was mostly ranked as the most important criteria, is crucial. Aspects around the user still need to be considered. E5 summarises: '*mostly, it is a trade-off between the accuracy and the comfort, that is what we realised — the more invasive, the more accurate in most cases*' (E5).

In general, a laboratory setting which would not benefit from apparent advantages of consumer goods, such as comfort, mobility, or a lower price, benefit from professional and clinically evaluated devices. These devices offer the benefit of being built for a professional use, with easier data exports. The advantages lie in the mostly more expensive price.

On the other hand are consumer wearables. Each of the expert interviewees have used consumer devices.[6] During the interview, the majority expressed their concerns regarding the data, i.e., they are '*accurate as you get from clinical sensors*' (E4). The mentioned reasons for still using these lay in their suitability for long-term or field studies (E3, E5, E1), additional insights for participants which are provided out-of-the-box (E3), to study '*concepts that work in the real world for a normal user*' (E2), and an increased comfort (E3).

Assessment of a Device. After the prioritisation of the design space dimension and the establishment of requirements, a wearable device can be assessed for its suitability.

The questions and guidance notes presented in Chapters 4 and 5 can support researchers in doing so. These guidelines mainly revolve around the data access, granularity and reliability. Assessing available documentations and related work can help researchers to gauge the suitability of a device. A good and clean manufacturer documentation of the device can help understand the format and granularity of the provided data. A rich body of scientific work where a device has been used can provide insights on how reliable data is. Especially clinical studies and the comparison to professional grade gold standard devices provides insights on the validity of the data. It is important that the conditions under which the data is collected are considered; data collection under free-living conditions can have an impact on validity of the data (Blythe et al., 2017; Dominick et al., 2016). The trust in a device and company can further be strengthened by it being widely available and well-documented; e.g., a proper documentation of how higher-level features have been calculated can support researchers in making assumptions on the validity.

But the impact of putting a wearable device on users should not be neglected. Especially during field studies and when researching participants in their everyday life, unobtrusiveness of the device are important. For assessing the comfort of a device, can be assessed on using standardised tools. E.g., (Knight et al., 2002) proposed a 6-item scale to assess the comfort when wearing wearable computers.

[6]For an overview of the used devices, see Appendix I, Table A.2

Lastly, the trust in a device company and a device itself should be considered. Start-ups and crowdfunded campaigns may offer the 'perfect' solution; still, 90% of startups fail [7]. But even large companies, such as Microsoft, can stop the support for a device, i.e., *Microsoft Band 2*. Investment in such a device, which is shortly after discontinued can lead to wasted funds.

6.5.3. Limitations

When conducting the analysis of the interview, it was found that some aspects and terms are subject to interpretation, e.g, the term accuracy had different definitions for different interviewees (see Section 6.3). Given the subjectivity that is always part of qualitative data, the assignment of the distinct sub-dimension to higher dimensions may be disputable for some cases. Likewise, the sample size of five expert interviewees is a small number; the aim of this study was not provide a comprehensive, in-depth assessment, but rather explore the space of researcher needs.

[7] `https://www.failory.com/blog/startup-failure-rate` (accessed: 01/07/2019)

Part V.

Application of the Design Space in an Exemplary Setting

Chapter 7

Applying a Wearable to Detect Interpersonal Synchrony during Conversational Interaction

This chapter presents a study on using wearable sensing to detect interpersonal synchrony using the Empatica E4 wristband wearable wristband during an unconstrained speed networking event. The design space developed in the last chapter has been applied to the E4 wristband device to evaluate its suitability. Preliminary results of this study have been published at the WearSys workshop 2018 (Hänsel et al., 2018a). Further analysis revealed that parameters of interpersonal synchrony could not conclusively be detected in the collected sensing data during conversational interaction.

Throughout time, researchers were interested in learning more about inter-personal interaction. Work focused in learning about how people interact and what influences the way we interact. Especially sensing can support this research by making automated and in-depth data collection easier and supersede older approaches with manual observation and annotation videos. One field of interest is the phenomenon of *interpersonal synchrony* which describes the higher similarity of various signals between two people during interaction; this increased similarity can be related to the relationship of people or the quality and nature of interaction. As shown in the background Section 2.4, interpersonal synchrony has been explored using wearable sensing approaches.

The study presented in this chapter explored the space by using wearable sensing data from a wearable to detect *interpersonal synchrony* during conversational interaction in a natural setup of a speed networking event. However, the main focus of the chapter is rather on the question RQ 4: *How can future devices be evaluated for their suitability to be applied in a certain research setting?*. This question is addressed by examining an exemplary device — the *E4 wristband* for its suitability to be applied in the above discussed research scenario. With this in mind, the design space criteria from the previous chapter are considered for the *E4 wristband*; following, it is evaluated for its suitability to be applied in this study.

The study was undertaken with 24 participants during a 45 minute social speed-networking event where participants were free to interact and start conversations with each other participant. The collected sensing data was then used to calculate pairwise features of *interpersonal synchrony* and compare these for each participant pairing during interaction and when they were not interacting with each other.

7.1. Study Aim and Hypotheses

This study aimed to investigate how wrist-worn wearables equipped with conventional sensors, such as *accelerometer*, *Photoplethysmography (PPG)* heart rate, skin temperature, and skin conductance sensors, can pick up synchronisations in bodily signals of interacting pairs of people. This experiment is planned to test the below-presented hypotheses in a natural, unconstrained setting with a larger (20 to 30) number of participants simultaneously.

As previously discussed, related work showed the concept of *interpersonal synchrony* in social interactions. With the main aim of the study to identify social interactions based on *interpersonal synchrony* detected through wearable sensors, the main hypothesis to test is the following:

Hypothesis 7-8 *There is an increase in synchrony features between pairs of interacting participants compared to pseudo-pairs of non-interacting participants.*

The mentioned synchrony features are the features discussed in the previous Section 2.4.2 in the form of cross-correlation and correlation lag.

(a) front

(b) back

Figure 7.1.: Overview of the *E4 wristband* and its sensors.

Further, and in line with the previous laboratory study presented in Section 5.2, it is hypothesised that the collected physiological data of each participant represents their affective state and mood in the form of, e.g., arousal and valence (Witvliet and Vrana, 2007; Salimpoor et al., 2009). It is hypothesised:

Hypothesis 7-9 *There is a relationship between subjective mood samples and physiological data at various times throughout the experiment session.*

7.2. Choice of Wearable Measurement Device

With the above study aim and hypotheses in mind, this section a) presents the *Empatica E4 wristband* device, b) shows how the design space from the previous chapter has been applied.

7.2.1. The *E4 wristband*

The *E4 wristband* by Empatica Inc. has been selected to be considered as a wrist-worn sensor for this study. It is a research-focused sensing device, which employs common sensors used in commercial wearable devices, i.e. *accelerometer* and optical *Photoplethysmography (PPG)* sensors. On top of these, it also comprises skin temperature/heat flux and *Galvanic Skin Response (GSR)* sensors. Figure 7.1 shows the device, its sensors and user interface elements. It is advertised as a device to monitor the *Autonomous Nervous System (ANS)* responses. As related work has shown, a synchrony in *ANS*, and *Sympathetic Nervous System (SNS)* in particular, responses can be observed when people interact (Vanutelli et al., 2017; Tourunen, 2017). *ANS* responses are highly related to changes in heart activity, sweat glands activation (detected with *GSR*),and skin surface temperature. Natarajan et al. (2016) have successfully used the predecessor of the *E4 wristband* in their research on detecting devisions of *ANS* responses using wearables.

7.2.2. Design Space Application

This section applies the 7 dimensional design space extensively discussed and evaluated in Chapter 6. Firstly, requirements are discussed and prioritised. Secondly, each dimension is discussed to argue the suitability of the *E4 wristband*.

7.2.2.1. Requirements and Priorities

The discussed study on *interpersonal synchrony* in a natural, unconstrained social mingling scenario (2 hours) with 20 to 30 participants poses the following requirements:

Very importantly, since the study setup is supposed to be natural and unconstrained, the requirement on a device is to have a high *Comfort*, to not be restrictive, and having a sufficient *Mobility* without wires and without the need for a complex setup.

Secondly, the *Reliability* and *Data Richness* should support the hypotheses. A device which incorporates all the required sensors (motion sensors and physiological sensors) is required. For physiological sensing, a sapling rate of 1Hz or higher would be sufficient. Motion sensing data should be higher than 32Hz.

Thirdly, there are no very specific requirements for the *Data Accessibility* apart from data being available in a processable format. No real-time or wireless data access is required.

Fourthly, the device should be readily available and there should be reports that it is functioning and *Trustworthy*.

Less important are the dimensions around *Operability*. The device should be operable by the researcher after sufficient training. Study participants are not required to operate the device.

7.2.2.2. Reliability

The reliability of the device to show the hypothesised changes is crucial. To evaluate the suitability of the *E4 wristband* for the purpose to detect *interpersonal synchrony*, a literature assessment was conducted.

The validity of the *E4 wristband* to detect changes in the *ANS*, e.g., stress responses, was supported by a rich body of research. In terms of detecting subtle changes, e.g., as changes from the *ANS*, the *E4 wristband* has shown to be reliable (Natarajan et al., 2016). Ragot et al. (2017) compared the *E4 wristband* with a laboratory device — Biopac MP150 — for two-dimensional emotion recognition.[1] The *E4 wristband*'s heart rate data showed a .99 correlation with the laboratory device. The accuracy for the emotion recognition was comparable between the two devices — around 70% for arousal and 66% for valence. Simmilarily, Gjoreski et al. (2017) used the *E4 wristband* successfully for binary stress classification. Pietilä et al. (2017) reported a good accuracy of the heart rate compared to the laboratory *Electrocardiogram (ECG)* device, as long as there is no no excessive wrist movement, i.e., during household work. The manufacturer themselves provided evaluations of the heart rate and *Electrodermal Activity (EDA)* sensor of the wristband showing high correlation to laboratory devices (Empatica Support, 2017b,a).

There are several works which rely on using the *E4 wristband* (or its predecessor the E3) device to detect *interpersonal synchrony*. Haataja et al. used the E3's *EDA* data to shows significant synchronisations amongst pairs of pupils in the classroom. In a similar setting, Di Lascio et al. (2018) showed that synchrony features from the *E4 wristband*'s *EDA* contribute to engagement classification in classrooms. Grafsgaard et al. (2018) used the E3's accelerometer data to model nonverbal synchrony in romantic couples. Ward et al. (2018) showed accelerometer synchronisation amongst autistic children wearing the *E4 wristband*. Gashi et al. (2019) investigated whether the *E4 wristband EDA* and heart rate synchrony strength is related to engagement between presenters and audience. They showed the relationship for *EDA* features but not for heart rate features. von Zimmermann et al. (2018) looked into movement data

[1] Ragot et al. (2017) administered the *Self-Assessment Manikin* but just used the valence and arousal dimension. For an overview of affect models, refer to Section 2.3.1

from the *E4 wristband* and the synchronisation amongst participants while they performed choreographic tasks; they successfully showed that coordination predicts the group affiliation.

Based on the above work, the *E4 wristband* was considered suitable in terms of reliability, and data reliability in particular, of the device.

7.2.2.3. Comfort of Attachment

This study investigates the phenomenon of *interpersonal synchrony* in a natural, unconstrained setup. The requirement for a suitable device is that it is easy to attach and not restricts the participants. Wrist devices are in general very wearable and comfortable; further, research has shown the high social comfort of wrist-devices Profita et al. (2013); Dunne et al. (2014). Additionally, some of the above discussed work using the *E4 wristband* were in natural settings, e.g., classrooms (Di Lascio et al., 2018) or when working with autistic children during a theatre performance (Ward et al., 2018). The *E4 wristband* is a wristband with a low effort of attaching it to the body and a low risk of causing discomfort. The adjustable strap makes it suitable for different wrist sizes.

7.2.2.4. Mobility

The requirements for a natural, unconstrained study setup require a device which is mobile, can be used without the need to be attached to a computer or power source for the duration of the study session (around 2 hours). Further, since the participant number is high (20 to 30 participants), the device needs to be very easy to setup.

The *E4 wristband* is a very mobile device with no cables to restrict mobility. It further has a long battery life of over 36 hours in recording mode (Empatica Inc., 2014); this is more than sufficient enough for the 2-hour experiment session.

7.2.2.5. Data Richness

To show the synchronisation of *ANS* and movement signals during social interaction, the requirement was for a device to provide this data. Especially movement data (von Zimmermann et al., 2018; Katevas et al., 2015) and physiological signals, i.e., *EDA* and heart rate, have shown to synchronise (Vanutelli et al., 2017; Karvonen et al., 2016; Mitkidis et al., 2015; Konvalinka et al., 2011).[2]

[2] An overview of *interpersonal synchrony* related signals is found in the background Section 2.4

The *E4 wristband* provides an extensive set of sensors included: heart rate, *EDA*, skin temperature, and accelerometer. The sampling rates of this data is high and between 4Hz (*EDA* and skin temperature), over 32 Hz (accelerometer), up to 64Hz (*PPG*) (Empatica Inc., 2014).

7.2.2.6. Data Accessibility

The data accessibility has to be easy and data should be obtained in a processable format, e.g., *Comma Separated Values (CSV)*. No real-time data transmission is needed for this study, since the data is analysed after the experiment.

The *E4 wristband* offers two modes of accessing the sensing data; the data can be downloaded when the device is connected to a PC or it can be real-time visualised on a companion app on a mobile phone. Since for this study setup, no real-time data transmission was necessary, and the first approach was followed.

7.2.2.7. Ease of Operation

The device used within this study should be easy and straight forward to use. The device was thoroughly tested before the study session. While the wristbands merely provide limited feedback in the form of an LED indicator[3], the handling during the study was easy enough. The event button is pressed for 1 to 3 seconds to activate and stop the recording. A shorter press places event markers. In retrospect, some data has been lost since participants pressed the button too long when instructed to place a marker resulting in the recording to stop prematurely.

7.2.2.8. Trustworthiness

The *E4 wristband* by Empatica is a reliable and established device used in several research settings, as previous related work showed. The company is an MIT spin-off, and it exists for many years with a range of evaluated devices. Since the device is especially marketed for research, the manufacturer support for these use cases is given.

[3]The LED indicator of the *E4 wristband* supports 6 different colors and differing blinking patterns to indicate errors or the device status, see `https://support.empatica.com/hc/en-us/articles/203381359-E4-Led-Guide` (accessed 15/02/2014)

7.2.3. Further considerations

The design space from Chapter 6, also contains the themes *Financial Cost*, *Social Acceptability*, and *Privacy*.

No personal data of the participants, e.g., demographics or identifying information, is shared with the manufacturer of the *E4 wristband* device.[4]

The social acceptability aspect of the device is not applicable. For this short-term study all participants are briefed about the device and no interactions which could have been perceived as 'weird' by others were required.

The aspect of financial cost is also not applicable in this study setup; the 30 *E4 wristbands* were kindly provided by Guido Orgs (Goldsmiths, University of London) and Daniel Richardson (UCL).

7.3. Study Design

The study is intended to detect *interpersonal synchrony* using the wearable *Empatica E4 wristband* in an unconstrained natural setting. The following sections describe the design of the study.

7.3.1. Setting

As one objective of the study was to collect data in a natural, reasonably unconstrained setting, the decision was made to set up a networking event. During this event, it was envisioned that participants come together in pairs (or groups) and talk about topics they choose. During the experiment wearable (and mobile) data was collected from each participant. While this thesis

[4] Privacy notice of Empapica: `https://support.empatica.com/hc/en-us/articles/202524239-What-does-Empatica-do-to-protect-end-user-privacy-` Accessed: 10/06/2019

focuses on evaluating the wearable device data, the following sections include the setup of mobile phones for the sake of completeness.

The experiment was split into two sessions: one session which was aimed to facilitate mainly one-to-one interactions to avoid interferences and complex dynamics from larger group interactions and a second session which allowed any pair or group formation. These sessions were aimed to last 30 minutes each[5].

7.3.2. Material

Apart from demographic information on each participant, relationship and closeness per participant pairing, as well as, personality and empathy of every single participant were assessed. An overview of the administered questionnaires can be found in the appendix in Table A.3.

Empathy, Demographics and Personality. Before the actual experiment, general demographic data on age, gender, and employment were collected. Furthermore, the Big Five personality traits were enquired using the *10-item Big Five Inventory (BFI-10)* by Rammstedt and John (2007). The empathy of participants will be assessed using the *EQ Scale* by Loewen et al. (2010). This is an 8-item scale to assess general empathy. These questionnaires were administered as an online questionnaire.

Relationship, Closeness and Sympathy. The relationship of participants will be assessed before and after the actual experiment session. A questionnaire containing neutral facial photos of all other participants is used to account for participants potentially not knowing each other; this leads to bi-directional pairings of responses, i.e., a relationship response from participant A to participant B and vice versa. The relationship is assessed by using free-text responses. The interpersonal closeness between participants is assessed using the pictorial *Inclusion of Others in Self (IOS) Scale* by Aron et al. (1992). It is a scale showing 10-levels of closeness via two circles with different overlap; once circle represents oneself and the other represents the opposing person. Perceived sympathy was assessed using a 10-point scale.

[5] During the actual experiment, due to delays, the second session had to be terminated early after 10 minutes to keep in schedule.

Interaction Recall. The interaction recall was assessed twice during the experiment sessions: after the first session of one-to-one interactions and the second session. Participants indicated their responses for each other participant using the neutral facial photos. It aided to assess if participants recalled an interaction and the remembered quality of the interaction.

Mood. Two common mood questionnaires were administered to trace changes in mood and affect throughout the experiment. The *Self-Assessment Manikin* by Bradley and Lang (1994) is a pictorial scale to capture perceived valence, arousal and dominance in subjects. The *Positive Negative Affect Schedule (PANAS)* by Watson et al. (1988) is a 20-item scale with affect words, e.g., interested, ashamed, enthusiastic. Participants rate each item on a 5-point scale to which extent they currently experience this emotion. Ratings are added to build two scores: positive and negative affect.

7.3.3. Participants and Procedure

The following sections describe the details of the participant recruitment, informed consent and study procedure[6]. 24 participants, 9 male and 15 female, took part in the actual experiment session. The material can be found in Appendix J.

7.3.3.1. Recruitment Participants were recruited locally at Queen Mary University of London via word of mouth, mailing lists and flyers; participants were reimbursed with £ 20 for their time. Participants were selected based on their mobile phone model and demographics to provide a balanced age and gender distribution. There were 8 iOS mobile phones available to be handed out; the rest of the participants used their personal mobile phone. Suitable participants were contacted and briefed about the further procedure.

7.3.3.2. Pre-study preparations Five days before the experiment session, participants were asked to install the *CrowdSense* mobile sensing application on their personal iOS device. In case that the participants have been selected to get handed a device by the research team, this step was omitted. Further, they were asked to provide a well-recognisable

[6]Although the focus of the work presented in this thesis focuses on wearable sensing with the *E4 wristband*, the descriptions include the setup of the mobile sensing (including Bluetooth beacons) for the sake of completeness.

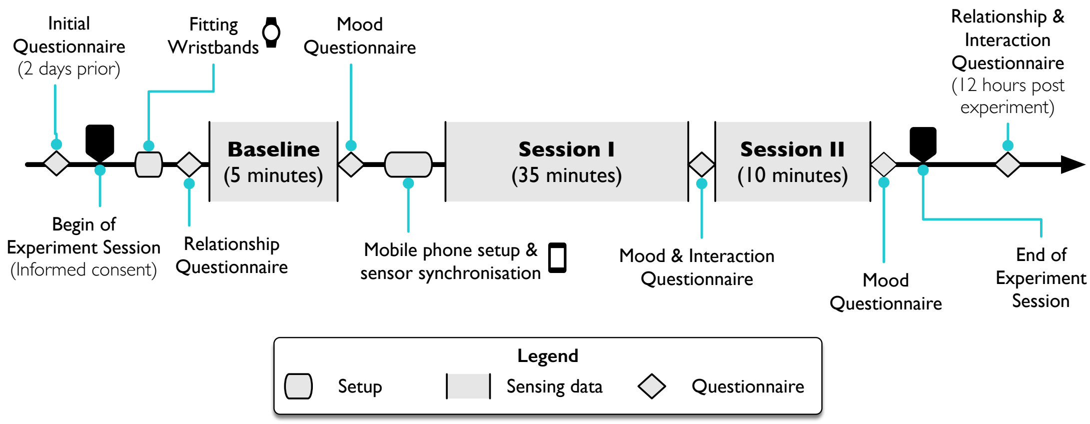


Figure 7.2.: Overview of the study procedure

facial photo which was used for creating relationship questionnaires. The demographic questionnaire assessing age, gender, ethnicity, Big Five personality traits using the *BFI-10*, and the *EQ Scale* was also administered as an online questionnaire.

All 24 participants, who provided photos and answered the pre-study questionnaires beforehand, showed up for the actual experiment.

7.3.3.3. Experimental Session

On the day of the experiment, participants were invited to the Performance Lab of Queen Mary University of London. The Performance Lab is an event space often used for social events. It is suitable for these experiments due to its isolation from environmental factors and external noise. The experiment space has a dimension of 6.57 × 5.36 m. The Performance Lab was equipped with two orthogonal facing cameras to record video (but not audio due to privacy reasons) during the experiment. The floor plan of this space and placements of the cameras can be found in the appendix in Figure Appendix K.

Introduction and Preparations. On arrival, participants were greeted, handed an information sheet and signed a consent form. They were then equipped with an *E4 wristband* and a Bluetooth beacon (Radius Networks RadBeacon Dot[7]) to put in their pocket — the beacon placement in

[7] https://radiusnetworks.com

the left or right pocket was randomly predetermined. After, participants were asked to fill out a relationship questionnaire to indicate their relationship to each other participant.

Baseline Recording. For recording a baseline of the physiological signals, participants were led in a quiet room to sit down for five minutes and relax. They filled out a mood questionnaire.

Sensor Setup and Synchronisation. After the baseline recording, participants were led into the experiment space in the Performance Lab. There, the mobile phone sensing app was set up collectively. For synchronising the mobile and wearable sensing sources with the video feed, participants performed a wave gesture in from of the cameras by holding their mobile phone in the wristband-equipped hand. This later allowed the synchronisation of accelerometer data across the devices and the video feed. After the setup, the phone was placed in the pocket without the Bluetooth beacon dot.

Session I. Then, the participants were instructed to interact in one-on-one iterations and change conversation partners frequently. This session (session I) lasted 35 minutes. In a short break, they filled out a mood questionnaire (*SAM* and *PANAS*) and an interaction questionnaire quickly assessing their recall on with whom they interacted and the quality of the interaction.

Session II. The following second session (session II) was shorter and lasted 10 minutes, but participants were unconstrained in the group size for interaction. Subsequently, they filled out a mood questionnaire. Due to technical difficulties with printing the final relationship questionnaire, this questionnaire was administered 12 hours after the sessions ended in the form of an online survey. It assessed the closeness to other participants using the *IOS Scale* and asked for the recall of interaction and the quality of interactions.

An overview of the steps to collect questionnaire and sensing data before, during, and after the experiment session is shown in Figure 7.2.

7.4. Data Collection and Analysis

The following sections go into the details about how the collected data (video, sensing, and questionnaire) was collected, processed and analysed. Finally, there is an overview of the data.

7.4.1. Ground Truth Data

Video footage was recorded from two different angles to collect ground truth data on the social interactions and group formations between participants. Two independent annotators marked the beginning of each interaction with a unique group ID using the ELAN software[8]. These annotations were cross-annotated and finally verified by a third person. The instructions for defining what interaction is was based on the definition by Kendon (1990):

> *An interaction begins at the moment two or more stationary people cooperate together to maintain a space between them to which they all have direct and exclusive access.*

These annotations were then exported from the software to be combined with the sensing data.

7.4.2. Sensing Data Collection

The raw collected sensing data from mobile phones and *Empatica E4 wristband* wristband was first collected and aggregated. The mobile phone data, which was collected using the *CrowdSense* app, was transferred from the participants' mobile phones automatically to the setup server; participants were matched to their assigned participant ID based the entered ID within the app and on a facial photograph taken in the CrowdSense app as a backup strategy. The data collection from the *E4 wristband* wristbands happened manually as each of them was connected to the PC and synchronised with the manufacturer's web dashboard; the data was downloaded from there, and the wristband ID was matched to the ID written down when every participant was handed their wearable. An overview of the sampling rates is shown in Table 7.1.

[8]ELAN is an open source video/audio annotation software `https://tla.mpi.nl/tools/tla-tools/elan/` (accessed: 20/01/2019)

		Sampling Rate	Unit
Physiological	heart rate	1 Hz	beats per minute (bpm)
	EDA	4 Hz	micro siemens
	skin temperature	4 Hz	°C
	Interbeat Interval (IBI)	event based	seconds (s)
	Blood volume pressure	64 Hz	— not documented —
Movement	3-axis acceleration (x,y,z)	32 Hz	1/64 G†

† gravitational force G (9.81ms^{-2} on the earth)

Table 7.1.: Overview of the sensor sampling rate and units of the *E4 wristband* wristband (Empatica Inc., 2014).

7.4.3. Preprocessing

The sensing data from the *E4 wristband* and the iOS mobile phones was preprocessed with the following steps:

1. Identification of missing data
2. Synchronisation of the data across devices
3. Cleaning and preparation of the data by identifying each experiment condition and exporting the data into separate files per condition, participant, and sensor source
4. Extraction of pairwise participant features per time window and ground truth labelling

The preprocessing was mainly performed using Python scripts.

Missing Data. During the scouting of the data, it became apparent that for some participants, the wristband did not fully record the whole experiment session, but merely around five minutes of data. For these participants, the data recording stopped after the baseline recording session. There, participants were instructed to shortly press the button on the wristband to place a marker. Pressing this button longer has stopped the data recording for 3 participants (P06, P12, P38). Oppositely, for one participant (P33), there is no baseline data, since she turned the wristband on after the baseline session.

Further, two participants (P02 and P27) terminated the experiment early and left after the first experiment session.

Data Synchronisation. As described before, the participants were instructed to collectively stand in front of the camera and perform a waving arm gesture for five times and once horizontally and vertically. This gesture was performed with their wristband-equipped arm and while they held their mobile phone in hand. The recorded signals were used for synchronising the sensor feeds with each other and the video feed. The accelerometer data from the *E4 wristband* and the mobile phone was synchronised using cross-correlation to find the most optimal time lag and manual visual inspection of the overlapped signals was performed. The video feed synchronisation was manually done for each participant.

Preparation and cleaning of the data. After the data was synchronised, it was split into separate files per experiment session, participant and sensor data source. There are unified timestamps across all the files. This timestamp matches the timestamp of the exported ground truth annotations on interactions from the video feeds. In the future, and after all final publications and analysis steps are completed, this dataset is planned to be released to the public.

Pairwise feature extraction and ground truth labelling. The final preprocessing step included the labelling of the data and extracted features based on the ground truth annotations.

The features were calculated per participant pairing and time window of 10 seconds. Apart from single participant features, such as mean values and standard deviation per time window, the following synchrony features were calculated for each pair of participants:

max ccf: The maximum of the calculated cross-correlation values between the two participants' time series data

mean ccf: The mean of the calculated cross-correlation values between the two participants' time series data

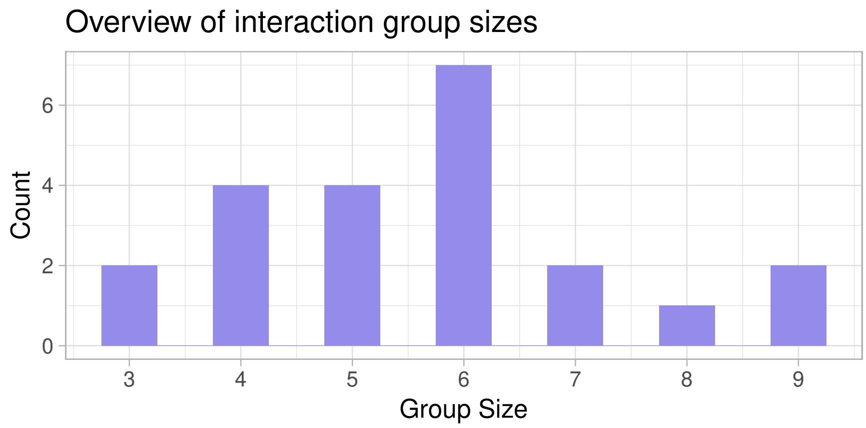


Figure 7.3.: Histogram showing the group sizes (larger than 2 people) throughout the experiment (mainly Session II). Group configurations lasting less than 5 seconds are omitted from this statistic.

lag of max ccf: The time shift (lag) of the maximum of the calculated cross-correlation values between the two participants' time series data

distance: The absolute difference between the means of the two participants' time series data

These calculated features were annotated with ground truth labels on whether the pair of participants was interacting during that point of time or not.

7.4.4. Tools of Analysis

The analysis of the data was performed using R. Statistical tests in the form of t-tests, correlation and nonparametric alternatives were performed using the `stats` package of R. Cross-Recurrence Quantification Analysis was performed using the `crqa` package.

7.4.5. Interaction and Pseudo Pairs

For the investigation of increased synchrony in interacting pairs of participants, surrogate data for comparison is needed (see Section 2.4.2). Since for every pair of interacting participants, there was also data of available for when they were not interacting with each other but someone else, this data can be used as surrogate data and in this case the non-interacting participants act as pseudo-interacting pairs (short pseudo pairs).

In total, 99 one-to-one interactions were observed with a mean duration of 254.9s (±161.7s) and 22 group interactions (i.e., interactions that include more than two participants) with a mean duration of 117.2s (±139.4s). A separate interaction begins when the members of a group change (e.g., someone joins or leaves the group). If the group configuration consisted of less than 5 seconds, then it is not counted in this statistic. An overview of the group interactions is depicted in Figure 7.3.

7.5. Results

While the previous sections describe the study setup as a whole (including the mobile phone sensing data), the results will focus on the data from the *Empatica E4 wristband*. Further, the detection of synchrony features is performed on the first session of the experiment; Session I focused on one-to-one interactions; this reduces the influence of complex factors in non-dyadic interactions and makes interpersonal synchrony detection harder. Due to the flowingly presented, non-conclusive results, the analysis methods have not been further investigated for Session II.

7.5.1. Subjective mood and sensor readings

Three mood questionnaires utilising the *Self-Assessment Manikin* and the *Positive Negative Affect Schedule (PANAS)* scales were administered; these were timed after the baseline recording and the first and second session. These mood ratings were compared and correlated to the average heart rate, *Electrodermal Activity (EDA)* and skin temperature readings in the last 5 minutes before the mood questionnaire was administered; the sensor readings were intra-person z-normalised to account for high variability of baseline physiologies between subjects (see Lykken et al. (1966)). There were no significant ($p < .05$) Pearson correlations observed (see Table 7.2); this

	SAM			PANAS	
	Valence	Arousal	Dominance	Positive Affect	Negative Affect
Heart Rate	.169 p=.245	.154 p=.291	-.011 p=.940	-.067 p=.662	-.048 p=.743
EDA	.021 p=.884	.120 p=.169	-.061 p=.676	-.026 p=.864	-.162 p=.266
Skin Temperature	.041 p=.782	-.068 p=.639	-.027 p=.856	-.025 p=.873	-.071 p=.628

Table 7.2.: Relationship of subjective mood ratings and sensor readings. Pearson correlation coefficients and significance signifiers for the mood ratings (collected 3 times during the experiment session using the *SAM* and *PANAS* scales) and sensor readings (averages of 5 minute window before the mood sample was collected). There were no significant relationships detected.

is contrary to the exception that mood and affective states are related to physiological sensor readings as it has been shown in related work (Witvliet and Vrana, 2007; Salimpoor et al., 2009).

7.5.2. Correlation Features

Interpersonal synchrony is related to a higher degree of synchrony features during interaction compared to non-interaction (see Hypothesis 7-8). Correlation and cross-correlation (with a time lag) are therefore expected to be higher during interaction compared to non-interacting pseudo pairs. The following sections present the results.

7.5.2.1. Accelerometer correlation As a first step, exemplary data snippets were explored and plotted to gain an understanding of the data. An exemplary plot of calculated, windowed Pearson correlations from P26 and P37 are shown in Figure 7.4. It

		mean (SD)		Mann-Whitney U)	
		interacting	non-interacting	U	p
Accelerometer Magnitude	max of ccf	0.984 (0.25)	0.991 (0.24)		n.s.
	mean of ccf	0.958 (0.24)	0.965 (0.23)		n.s.
	lag of max ccf	2.462 (0.83)	2.448 (0.84)		n.s.
	distance	0.106 (0.224)	0.097 (0.218)	32198964	.001

(a) Movement synchrony features

		mean (SD)		Mann-Whitney U)	
		interacting	non-interacting	U	p
Skin Cond.	max of ccf	1.43 (4.73)	0.93 (2.44)		n.s.
	mean of ccf	1.35 (4.54)	0.86 (2.28)		n.s.
	lag of max ccf	2.91 (0.93)	2.94 (0.06)		n.s.
	distance	1.13 (1.37)	1.27(1.47)	28740000	.000
Heart Rate	max of ccf	7594.79 (1743.21)	7588.93 (1627.32)		n.s.
	mean of ccf	7511.28 (1720.1)	7505.16 (1604.56)		n.s.
	lag max ccf	3.4 (0.61)	3.39 (0.63)		n.s.
	distance	13.56 (12.48)	14.23 (12.36)	29230000	.002
Skin Temp.	max of ccf	1170.01 (123.36)	1178.35 (130.62)	29280000	.003
	mean of ccf	1166.39 (121.15)	1174.78 (128.97)	29276000	.003
	lag max ccf	3.00 (0.88).	3.03 (0.89)		n.s.
	distance	3.31 (2.47)	3.22 (2.24)		n.s.

(b) Physiological synchrony features

Table 7.3.: Mean and standard deviation of wearable sensor features during interaction and non-interaction. Features include the max and mean of the cross-correlation function (ccf), the lag (shift) of the maximum calculated cff and the mean absolute distance of the measures between the participant pairs' data. (n.s. – not significant)

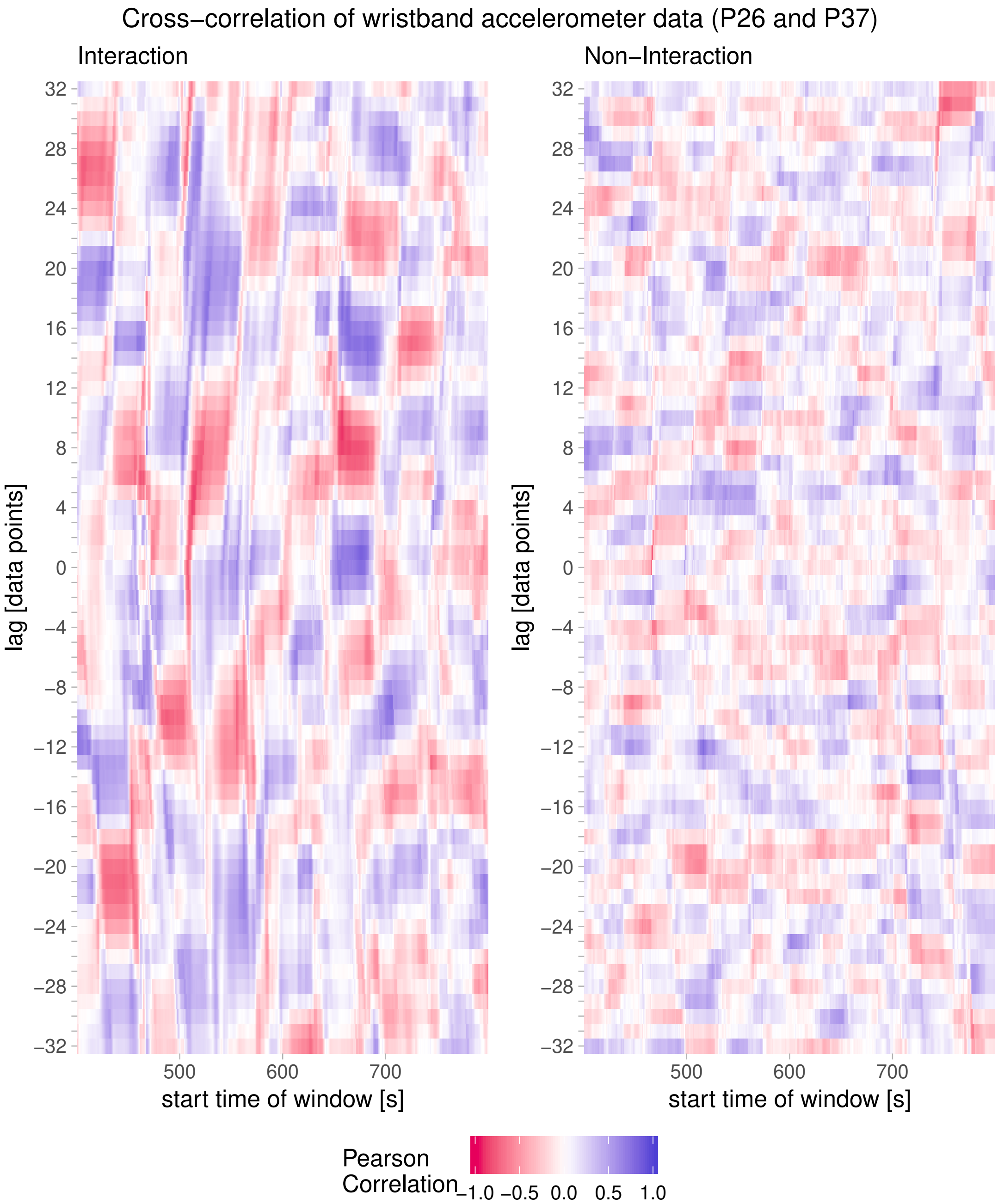


Figure 7.4.: Exemplary plot of the Pearson cross-correlation of *E4 wristband* accelerometer data of P26 and P37 with a maximum lag of 1 second (32 data points) in each direction.

shows the Pearson correlation coefficients for the *E4 wristband*'s accelerometer magnitude[9] data in a time window of 1 second with a sliding window of 1 second and a window increment of 5 data points (sampling rate was 32 Hz). The maximum lag of data points is 1 second in each direction (which is equivalent to 32 data points, since the *E4 wristband*'s accelerometer sampling rate is 32 samples per second). This method has been adapted from Boker et al. (2002). The non-interaction plot includes data where P26 and P37 were not interacting with each other, but other participants[10]. It is apparent that the correlation coefficients during non-interaction (right-hand-side plot) appear more shuffled and random than the correlation coefficients during interaction (left-hand-side). This is also reflected in the mean absolute correlation strength which is higher during interaction ($\overline{|\rho|} = 0.230$) compared to non-interaction($\overline{|\rho|} = 0.168$).

It has to be noted that this is just an exemplary snippet and not representative of the whole data set. The tendency of how accelerometer data correlated over all participants and collected data points was investigated using correlation synchrony features per time window and participant pairing. Statistical comparisons in form of *Mann–Whitney U tests* were performed[11]; the results are presented in Table 7.3a.

It was expected that the maximum strength of the cross-correlated accelerometer magnitude would be higher in interacting participant pairs than non-interacting pseudo pairs; this was not confirmed. The mean values show a non-significantly higher cross-correlation maximum during non-interaction. Similarly, the mean strength of cross-correlations per time window was expected to be higher during the interaction; this was also not the case.

The lag feature is the time shift of the data series to get the maximum cross-correlation strength. The higher the lag, the more the time series have to be shifted. It was expected that the lag is lower during interactions; this was also not the case.

Lastly, considering the mean distance between two participants' accelerometer readings showed a significant higher distance of interacting participants compared to non-interacting pseudo

[9] The accelerometer magnitude is the square root of the sum of squared x, y, z components. The accelerometer magnitude shows the directionless strength of the acceleration.

[10] P26 interacted with P05 and P37 interacted with P19

[11] The normal-distribution of the synchrony features was tested using *Shapiro–Wilk tests* and visual inspection of histograms and QQ-plots

pairs; this again is contrary to what would have been expected. The signal of two interacting participants was expected to be more similar, ergo resulting in a lower distance.

Overall, the results from considering the accelerometer data of all participant pairings are inconclusive or even contrary to the expectation. The increased similarity of the accelerometer and movement signals cannot be shown, and Hypothesis 7-8 could not be confirmed for movement synchrony features.

7.5.2.2. Physiological Singals

Similar to the movement data in the previous section, physiological signals features were calculated in 10-second windows. The mean and standard deviation of these features are presented in Table 7.3b. It was expected, that maximum and mean cross-correlation are higher in interacting than non-interacting pseudo pairs. While this is true for *EDA* and heart rate, the differences were not significant. There were significant differences in skin temperature readings but these were in the opposite direction.

The time lag between the maximum cross-correlation points was expected to be lower during interactions; this was true for *EDA* and skin temperature although not to a significant extent.

The distance (absolute difference) between the mean values of participants was expected to be lower and more similar during interactions. This effect was observed for *EDA* and heart rate and the differences were significant. On the contrary, the distance between skin temperature readings was higher during interaction (not significant).

In general, it can be said that the physiological differences between interacting and non-interacting pairs were very marginal; while the independent non-parametric t-test (*Mann–Whitney U test*) showed significant differences for some of the features, these may be caused by the considerable size of compared time windows (n = 43050). The results of this section do not confirm the Hypothesis 7-8 for physiological synchrony features.

7.6. Discussion

The following sections discuss the results from above and conclude with an outlook for future work.

7.6.1. Suitability of the E4

As discussed in this chapter, in Section 7.2.2, the *Empatica E4 wristband* device was found suitable for this study. The evaluation showed, that data was successfully collected from the device; merely data from 3 participants was missing due to operational errors: the wristband was accidentally turned off by participants when they pressed the device button. The data was suitable to be processed and analysed. However, the hypothesised effects were not detectable. The potential reasons are below.

7.6.2. Relationship of sensor readings and mood ratings

There was no correlation between the self-reported mood and the heart rate, *Electrodermal Activity (EDA)* and skin temperature sensor readings; this is contrary to the expectations and contrary to what has been observed in the previous studies within this thesis (see *EmoRate* study in Section 5.1 and the laboratory study in Section 5.2). One explanation could be that the sample was too small or that possibly the administration of paper questionnaires handed out by helpers influenced the mood ratings[12].

[12]The administration questionnaires were pre-labelled with the participant identifiers; this resulted in participants needing to find the right file for them to fill out and could have biased their mood ratings. In comparison, the mood ratings within the *EmoRate* app (Section 5.1) happened very unobtrusively and quickly on the *Apple Watch*. During the laboratory study of Section 5.2, participants were in a quiet, isolated space and the mood questionnaires were placed on a table.

7.6.3. Absence of detected synchrony

While it was hypothesised that interpersonal synchrony features were more prevalent during social interaction between pairs of participants compared to non-interacting pseudo pairs, this could not be confirmed for the collected dataset. The following sections discuss potential reasons for this.

Interferences. The experiment took place in the performance lab; this is a space isolated from external factors such as noise or light changes. The space was limited in size ($\sim 35m^2$) for the 24 participants. Neighbouring conversations or participants moving through the crowd to find another conversation partner could have led to interferences preventing interpersonal synchrony to be built up to a detectable extent. Related work from Slovák et al. (2014) discovered synchronisations in *Galvanic Skin Response (GSR)* during dyadic interaction in a natural setting (a pub) and outside of the lab, but still, participants were secluded from other non-participants. Further, participants were seated.

Task low in joint action. The task of socially conversing while standing in a room is very low in movements apart from personal gesturing. Many studies on interpersonal synchrony of movements focused on joint action tasks, like, e.g., dancing (Tarr et al., 2015; Reddish et al., 2013), walking (Katevas et al., 2015) or rhythm synchrony during collaborative music making (Morgan et al., 2015). A common finding was that increased synchrony is related to social bonding (Tarr et al., 2015) and cooperation.

Turn-taking behaviour during conversational interaction. Further, there is evidence that regular conversations lead to turn taking in gesturing happens; this means that while the speaker uses gestures often of the arms (Holler, 2010), listeners stand still with fewer arm movements. These observations could explain, while for accelerometer readings from the wristbands, there was less synchrony during the conversation. It is a field to explore in future work if wearables can be used to detect speaker and listener dynamics during dyadic social interaction.

Low affective task. Similarly, synchrony of physiological features has been often researched under exceptional circumstances. The famous fire-walk study by Konvalinka et al. (2011) found synchronisations between heart rate of active participants in the ritual and spectators related to them. The ritual of fire-walking over hot coals in itself is an extreme ritual leading to intense affective responses (Alcorta and Sosis, 2005). These high affective responses are easier to detect in physiological sensing signals, and therefore synchronisation of these intensified signals may be more prevalent and easier to identify. The task of conversing in a standard social setting, as the participants of this study performed it, may not lead to high enough affective responses and physiological signal changes.

No special participant relationship. The participants of this study were not selected based on their relationship with each other. Other studies specially recruited befriended participant pairs, e.g., Slovák et al. (2014), or considered special dyads, e.g., mother and child (Lunkenheimer et al., 2018; de Mendonça et al., 2010) or couples (Julien et al., 2000). Additionally, work has shown that pairs of people related to each other show higher synchrony (Konvalinka et al., 2011).

Reliability of the Device. The reliability of the *E4 wristband* device and the collected sensing data could be considered as a reason for the missing synchrony. However, the *E4 wristband* has been used in various previous research studies showing its validity compared to gold standard devices and the validity to detect *interpersonal synchrony* in various cases (see Section 7.2.2.2).

7.6.4. Conclusions and Future Work

The *E4 wristband* device was validates using the previously presented Design Space from Chapter 6. It was found to be a suitable device for this study. However, it was not possible to confirm that wearable sensor data from the *E4 wristband* can be used to detect increased interpersonal synchrony between interacting people in this unconstrained, natural setting of a networking event.

The potential reasons for this have been discussed above. These findings lead to the question if, in general, wearables are unsuitable for detecting interpersonal synchrony. While in more controlled setups, during joint action tasks (e.g., walking, dancing, etc.), or in highly affect-inducing settings synchrony has been detected, detecting synchrony in everyday life settings is challenging due to external factors or low physiological changes. Future work is necessary to explore this space and investigate how external biases and context can be detected and reduced.

Also, more advanced data filtering mechanisms can reduce noise in the data to identify subtle synchronisations in everyday life scenarios.

Overcoming these obstacles can lead to ubiquitous, in-the-wild research on how interpersonal interaction changes our physiological responses and how influencing factors such as relationship, trust, quality of interaction or engagement can be inferred. Gaining knowledge about these parameters in everyday life settings can not just enrich our understanding of how people interact but also be the stepping stone for technologies helping to promote better social interactions and support social connectedness.

Part VI.

Conclusion and Future Work

Chapter 8

Conclusion

Within this thesis, the potential of wearable physiological sensing was explored to address the question: *How suitable are physiology-sensing wearables for human-subject studies in affective and psychology research?* This question is not easy to answer; similar to the mobile device market, the wearable device landscape is highly fragmented with multiple vendors, proprietary systems, novel devices flooding the market, and little standardisation (Han et al., 2012). There is a low incentive or regulation around validating devices and their provided data. Loopholes in legislation for 'wellbeing' or 'fitness' devices allow devices to enter the market with a low threshold. While this promotes diversity of devices for differing consumer (and researcher) needs, it also enables unvalidated devices to flood the market. Researchers face the rough choice of what devices are applicable and suitable for being used in research.

Within this thesis, four exemplary devices were explored regarding their data access and suitability to be applied in affective research. Consequently, important choice criteria were derived and evaluated with semi-structured expert interviews. Finally, these were applied to a novel device. Below is a detailed summary of the research questions and what this work contributes to address these.

8.1. Summary of contributions

The following paragraphs summarise and reflect on the research questions and respective contributions of this thesis.

Research Question 1: *How can research-suitable data be obtained from wearable devices?*

This question was addressed in Chapter 4 by firstly considering and analysing four exemplary devices with different data access means. The different sensing data granularities and approaches how vendors provide data access in this devices has been shown and two implementational artefacts have been presented: *AWSense* framework (Hänsel et al., 2017) and LabExperiment app. Both artefacts are openly available for researchers to use. Following more general access strategies implemented in current, and prospective devices, are discussed. The main contribution of this chapter lies in summarising the findings and experiences around data access in wearables in the space of research applications. Consumer systems are often closed and do not grant access to raw sensing data. Proprietary processing means to derive the provided features, e.g., heart rate from optical *Photoplethysmography (PPG)* or burned calories, are not disclosed; this makes it difficult to gauge the appropriateness of the measures. It is discussed that while common access strategies exist, e.g., web *Application Programming Interfaces (APIs)* for accessing cloud stored data, these are subject to vendors supporting these. There is little standardisation on a) data formats or b) data access means (de Arriba-Pérez et al., 2016). The standards that exist, e.g., *Bluetooth Low Energy (BLE)* heart rate profiles, are just supported by selected device vendors. More work on developing and incentivising the support of standards, can make the integration of wearables in research easier and more long-lived. Considering data access in research, this leads to issues for researchers to consider, such as granularity of the data, privacy and ethical handling of participant data, and implementational efforts required. A list of aspects researchers need to consider when considering data access from a device has been discussed in this chapter.

Research Question 2: *How suitable are current state-of-the-art wearable devices to be applied for measuring stress and affect in the lab and-in-the wild?*

In Chapter 5, two research probes have been conducted to address the above research question. Firstly, the *Apple Watch* smartwatch was applied to show a relationship between sensing data and affect mood samples in the wild. Secondly, the four exemplary test devices were applied in the lab to show physiological stress changes. The lessons learned from both studies were summarised and discussed. The hypothesised observations were just partly confirmed within the two studies; especially the consumer devices were not able to show the hypothesised stress responses. Additionally and under movement, the consumer devices are less reliable than the professional device. Apart from the shown issues with data reliabilities, the contribution of this chapter lies in the discussion of experiences collected while conducting those studies. Properties of the devices, such as battery charging cycles or discomfort they caused, e.g., when participants removed electrodes them, are concerns to be considered. These findings are discussed and reflected upon. They contribute to the development of the design space in the next chapter.

Research Question 3: *What are criteria for choosing a suitable wearable sensing devices in a research settings?*

Chapter 6, firstly, reflected on the findings of the previous two chapters to extract five criteria to evaluate wearable physiological devices for their suitability. Aspects around the data, such as accessibility, richness, and reliability, exist as well as more user and convenience focused aspects, e.g., comfort and mobility. This *Initial Design Space (IDS)* was applied to the four test devices devices. Each of the devices has shown strengths and weaknesses in some of the dimensions. This highlights tradeoffs and compromises to be made when devices are applied in research settings. Following, this *IDS* design space was evaluated with five expert interviews. While the five initial dimensions were confirmed and mentioned by the experts, additional dimensions, i.e., Trustworthiness and Operability, and sub-dimensions were derived. This forms a more complete design space of using wearable physiological sensing in research. It is the core contribution of the presented thesis. These extracted dimensions can aid researchers to consider a wearable device from various angles.

Research Question 4: *How can future devices be evaluated for their suitability to be applied in a certain research setting?*

Chapter 7 uses the design space presented in the previous chapter and applies it to a novel device. It is discussed and how the chosen device, the *Empatica E4 wristband*, was suitable for the proposed research study on *interpersonal synchrony* detection. This chapter contributes in a) demonstrating how the design space can be applied, and b) providing a discussion on what was learned during applying the selected device in the research setup.

8.2. Future Directions

As highlighted in the related work chapter, the evaluation efforts of wearables mainly addressed issues around reliability and validity of the data. Numerous studies, especially from the field of medicine, compared devices under different conditions, e.g., physical activity. Other work focused on evaluating the devices in terms of user needs, e.g., Karahanoğlu and Erbuğ (2011). Little work is done in investigating other appropriateness measures of wearable devices for research. While this work provides the first steps into providing a design space with important factors determining a devices' suitability for a research setup, more work is needed.

What was learned from the experiences in conducting the studies within this these, but also from the expert interviews: The appropriateness of data from wearables is crucial. The data and its processing/analysis is what informs the novel knowledge in research; it is used to test hypothesis or deduct new theories. Therefore, the data provided by a device needs to be accessible, useable, and reliable. Future work on ensuring the successful use of wearables in research should address the issues of reliability and validity. While there is work on filtering of raw signals to increase validity, e.g., Wood and Asada (2007), many consumer devices do not offer this raw data. With novel devices continuously entering the market, easy and quick validation strategies are needed to establish the validity of sensing data under different conditions. In the interviews, participants expressed their 'feeling' that devices are unreliable (Section 6.3). Standardisations and regulations of this process could even be adapted by manufacturers to develop trust in their products and the provided data. Similarly, regulations and standards on how to access the collected sensing data in an ethical and privacy-preserving way can benefit all stakeholders involved. Research into the technical, organisational and ethical realisation is needed.

The research community should be more open and communicative on actual experiences gained while working with wearable devices. This effort should go beyond discussing the data. Many other aspects are equally important to consider. Participant-focused aspects, such as the comfort, or usability of a device for researcher and user alike are important. Things experienced during the usage of (not just wearable) tools in research can help peers and colleagues to not make more informed decisions.[1]

Concluding, the potential for wearables in research settings is immense; the yet new data streams can potentially allow insights into how humans behave, feel, and fare in the wild and outside the lab. But future work needs to address the challenges on how wearables can be easily, seamlessly and ethically be integrated into research processes.

[1] Also from an oeconomic and sustainability standpoint: Research funding should not be invested in acquiring devices which are then not used because it turned out they are not the right kind of device for a study. From the thesis' authors experience, once funding is available it is sometimes spend on hasty on purchases, compared to, e.g., a consumer spending his/her own funds.

Part VII.

Appendix and Bibliography

Appendix A

AWSense – Architecture and Demos

The *AWSense* libraries have been developed using Swift and are open-source available[2]. The architecture of the *AWSense* framework is divided in sub-modules: *Core* and *Connect*. An overview of the subcomponents is depicted in Figure A.1.

AWSense Core. The *AWSense* core module is a standalone library for the *Apple Watch*. It offers sensing functionality for the watch to ease access to the embedded sensors. It supports the following sensors from the *Apple Watch*:

- Heart rate
- Raw acceleration up to 100Hz
- computed device motion (linear acceleration, rotation rate, attitude, gravity) up to 100Hz

The core library can be included in an *Apple Watch* app. It offers a singleton `AWSensorManager` class to configure sensors, to subscribe to sensor updates, and to start/stop sensing sessions.

The sensor configuration allows for setting the frequency of the accelerometer and device motion data with a maximum sampling rate of 100Hz; the sampling rate of heart rate readings cannot be modified due to limitations by Apple. Within the framework, events are triggered when new data is available. The core library is meant for using sensing data within *Apple Watch* apps without the need for transferring the data to the phone. It could be used within, e.g., motion-enabled games on the wearable.

[2] AWSense GitHub repository: `https://github.com/MiezelKat/AWSense`

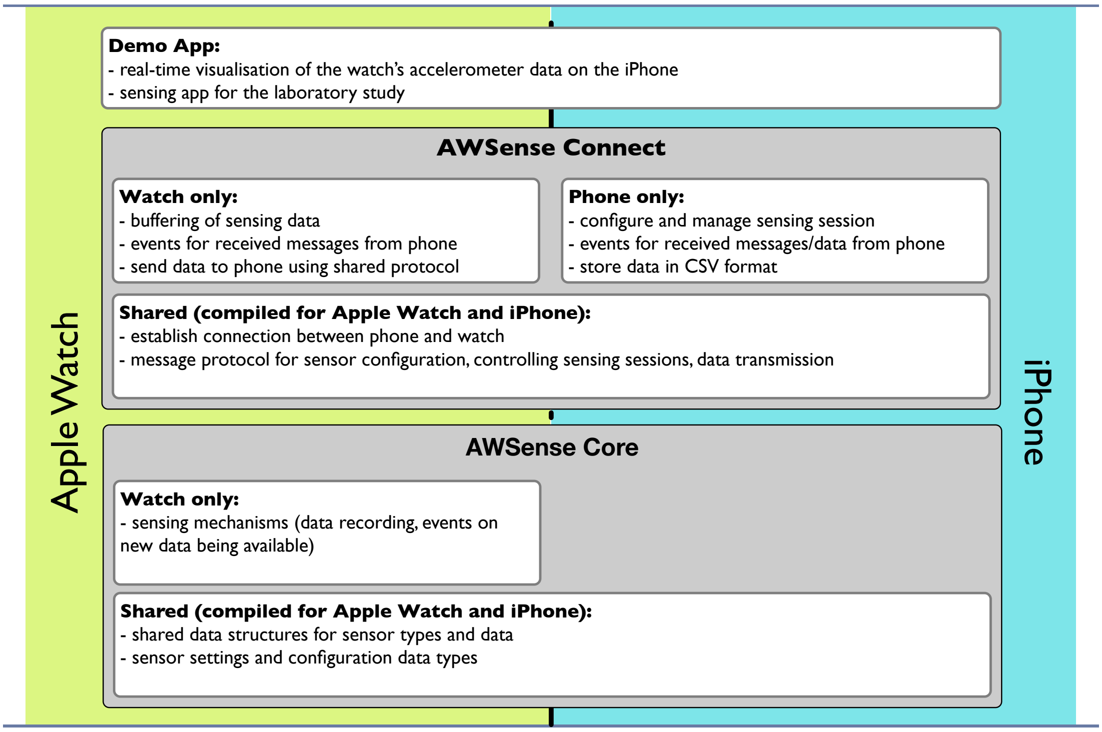


Figure A.1.: Architecture of the *AWSense* framework and its subcomponents — core and connect.

AWSense Connect. The connect module builds on top of the core module. It offers functionality for transmitting sensing data between the *Apple Watch* and iPhone; therefore it is available in a compiled form for iPhone and *Apple Watch*. The connect module, furthermore, enables the sensor configuration, event subscription and sensing session control from the phone, i.e., sensing sessions can be configured, started and stopped from the phone. A custom messaging protocol was developed to transfer sensing data and control messages between watch and phone.

The phone part of the connect module further contains a `RemoteSensingDataBuffer` singleton class for temporarily storing, as well as, persistently serialising the data in *Comma Separated Values (CSV)* format. These *CSV* files can then be downloaded from the iPhone.

Demo Applications

The functionality of the *AWSense* framework was demonstrated with three demo apps which were presented at MobiSys 2017 (Hänsel et al., 2017). These apps are part of the open-source project code from Github[3]. The framework further has been utilised for collecting sensing samples in the *LabExperiment* app (see Appendix B) during the laboratory experiment on evaluating the *Apple Watch* and two other consumer devices.

AWSense Core Demo. The watch-only demo app is run on the *Apple Watch* to demonstrate the *core* part of the *AWSense* framework. It just includes the watch-side library for easy access to the sensors. The demo app shows the sensing data (raw acceleration and linear acceleration from device motion) in real-time on the watch. An overview of the interface is depicted in Figure A.2.

AWSense Connect Demo. This watch and phone app is demonstrating the overall functionality of the *AWSense* framework and the *core* and *connect* libraries. It offers an interface on the iPhone to configure a sensing session by selecting the desired sensor sources, transmission interval for the messages, and the session name (which is encoded in the exported *Comma Separated Values (CSV)* files). Once a sensing session is started, it shows the first data sample of the last received message. An overview of the interface is shown in Figure A.3.

AWSense MobiSys Demo. This watch and phone demo app is demonstrating the capability of near real-time data transmission from the watch to the phone[4]. The phone app starts a sensing session for getting device motion readings with 3Hz and a message transmission time of 0.2 seconds. The received gravity data from the device motion is converted to show the tilt of the watch and is presented on a 2D plane on the phone (cf, Figure A.4).

[3]Github AWSense: `https://github.com/MiezelKat/AWSense`
[4]It is noted that the data transmission is near real-time since internal processes out of control can delay the delivery of messages front the watch to the phone.

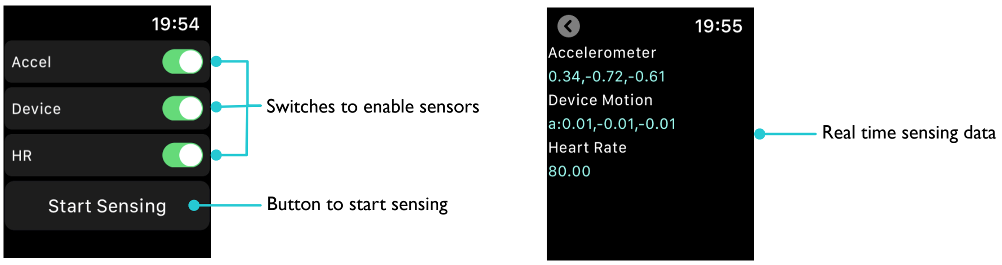


Figure A.2.: AWSense core demo application screenshots presenting the watch interface to configure and start a sensing session (left-hand-side screenshot) and the real time visualisation of sensing data (right-hand-side screenshot)

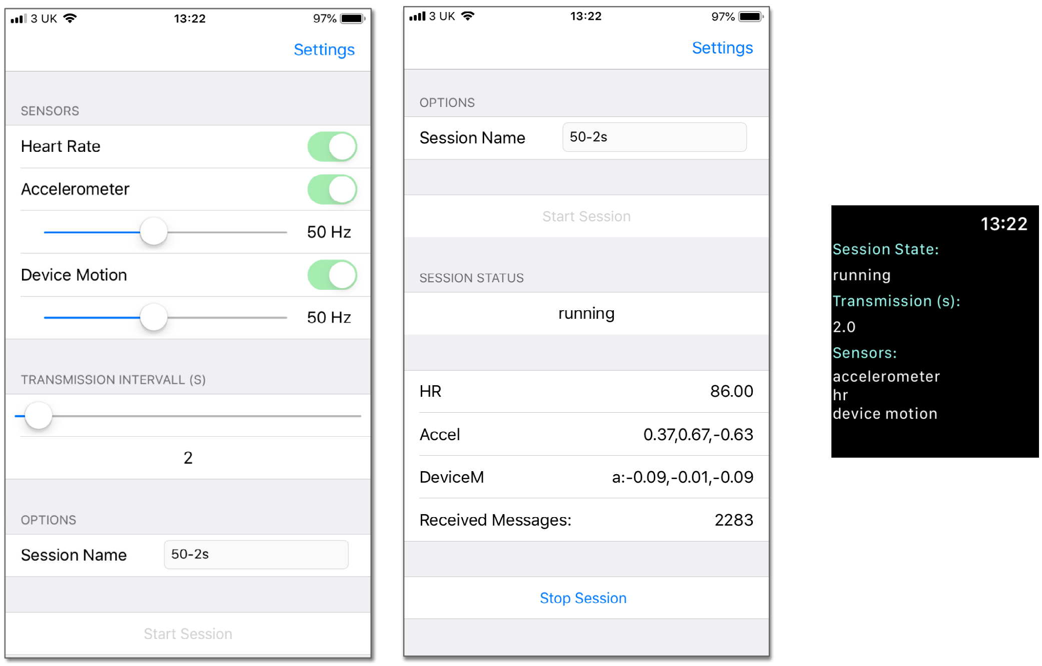


Figure A.3.: AWSense Connect demo application screenshots. The left-hand side screen shots show the phone interface to configure (selection of sensors, sampling rates, data transmission interval), start, and stop a sensing session. The right-hand side shows the Apple Watch interface.

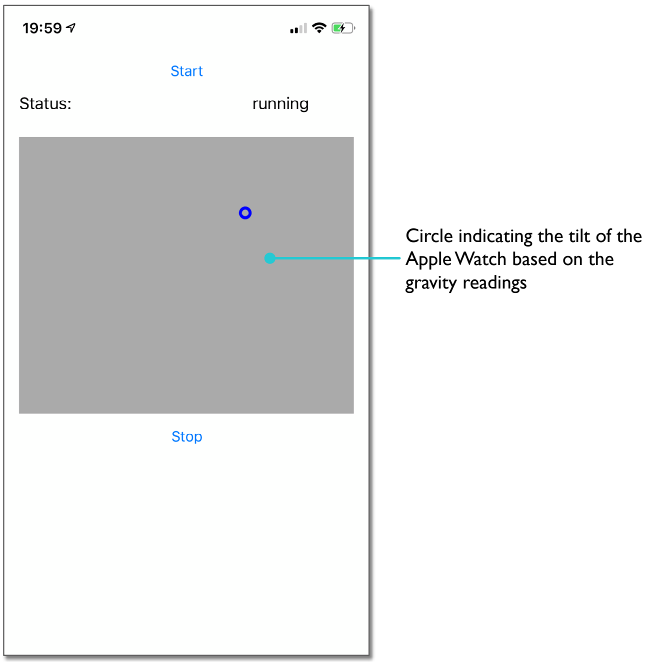


Figure A.4.: This demo application presents the real-time data transmission capabilities of *AWSense*. The phone app presents the real-time sensing data from the *Apple Watch* by plotting the tilt of the watch (derived from the gravity device motion data) on a 2D plane.

Appendix B

LabExperiment App – Architecture and UI

Below, there are Figures of an architectural overview and *user interface (UI)* screenshots of the LabExperiment[5] app presented in Section 4.2.3 and used in the laboratory experiment (Section 5.2).

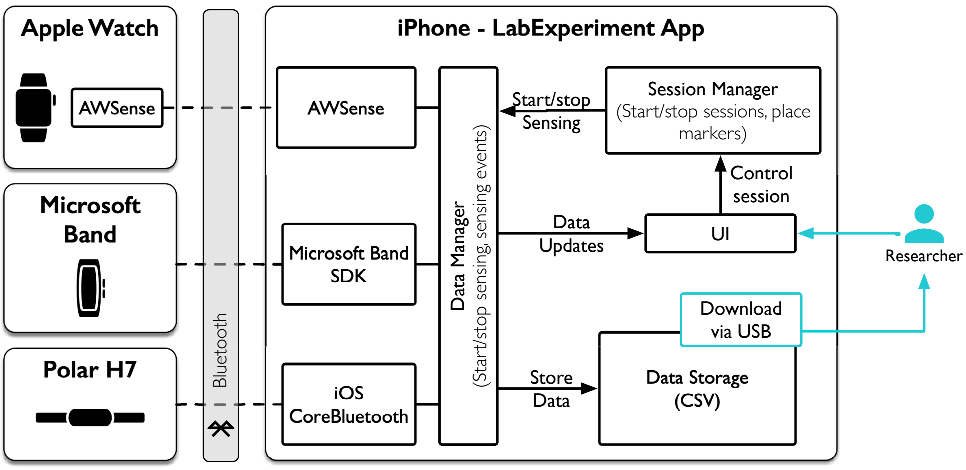


Figure A.5.: Architecture of the LabExperiment app for collecting data from *Apple Watch*, *Microsoft Band 2* and *Polar H7* for applications in laboratory experiments.

[5] source code available on Github `https://github.com/MiezelKat/LabExperimentApp`

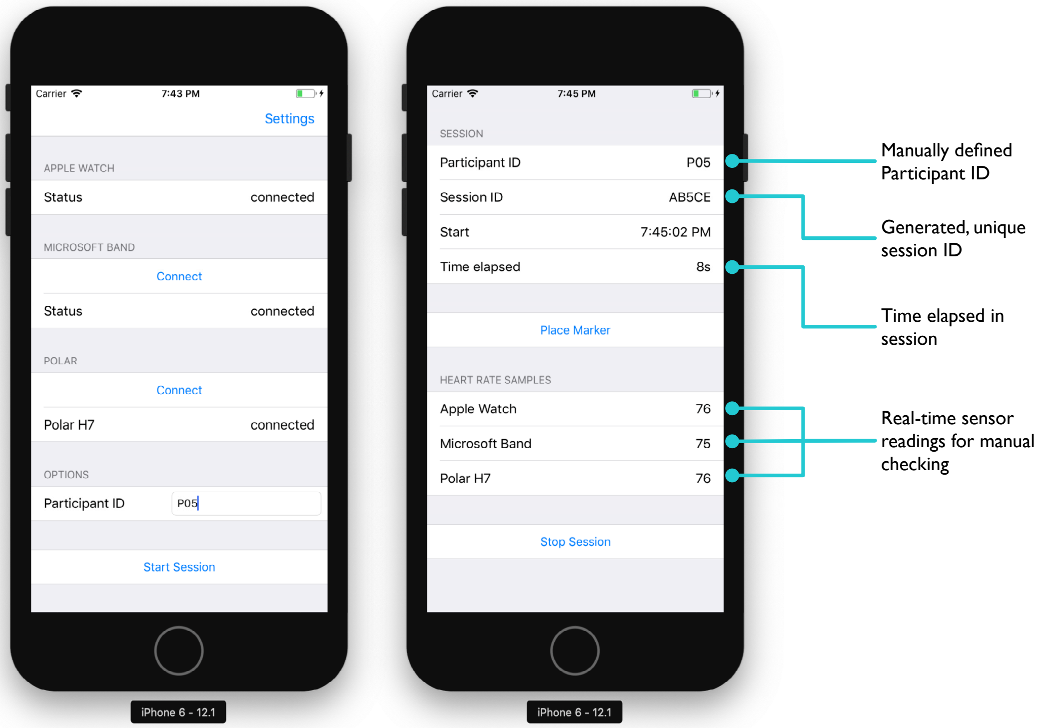


Figure A.6.: Screenshots of the LabExperiment app for collecting data from *Apple Watch*, *Microsoft Band* and *Polar H7*.

Appendix C

EmoRate – App UI

Apple Watch App. One part of the *EmoRate* application is the *Apple Watch* app; beyond the sensing data collection, its purpose is to allow users to enter their current emotional state quickly. The *Apple Watch* provides an easy and straightforward interface.

Figure A.7.: Schematics of the *Apple Watch* User Interface

The *Apple Watch user interface (UI)* for daily emotional self-assessments is designed in a straightforward way. It should allow the user to quickly enter the current emotional state within 5 to 20 seconds. The user is supposed to choose one of five valence options and one of five arousal states to describe his current affective state and mood. For each of these states, there is a separate page, where the user can select one of the five choices by using the 'Digital Crown' of the watch; this enables an easy and quick selection. By swiping left and right, the user can change between pages for the emotional state, arousal state, and a summary page, which allows the submission of the assessment. The summary page also allows disabling location logging for the current assessment for privacy reasons. Furthermore, there is a help screen to explain the interface of the app to the user. An overview of the app pages is shown in Figure A.9b.

Phone App Pendant. The Phone App has mainly two functions; it asks the user to fill out the more complex questionnaires and assessments, and it stores the assessment results and presents them to the user to reflect. Figure A.8 shows how the consent form is presented on the watch.

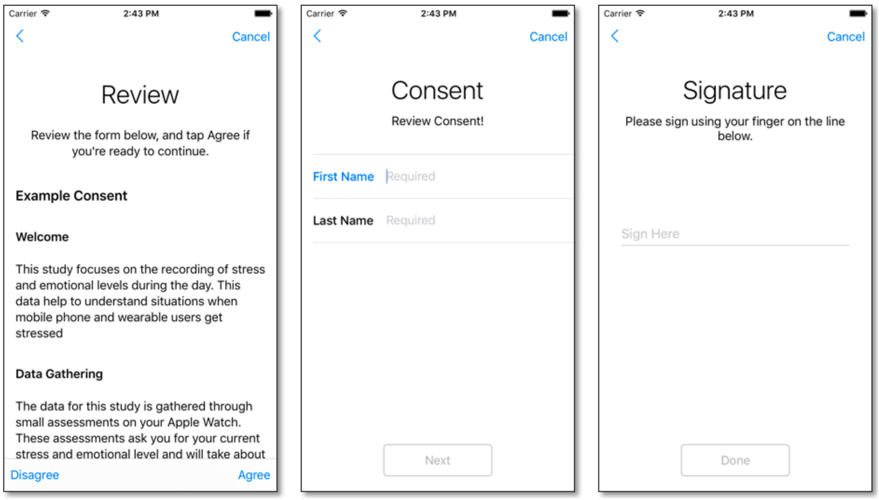
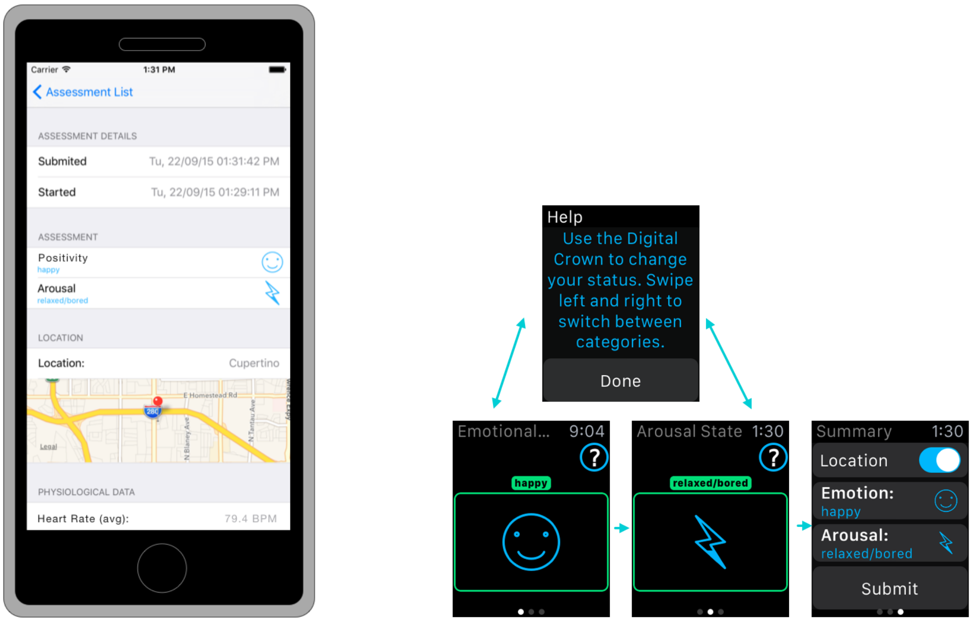


Figure A.8.: Snippets from the Informed Consent of the *EmoRate* App

Figure A.9a shows a summary of one of the daily assessments, this includes information on the time the assessment has been taken, the selected emotional and stress states, and the location; and it supports the reflection on emotional states in the past.



(a) iPhone app

(b) Apple watch app

Figure A.9.: *UI* of the *EmoRate* phone and watch app. The phone pendant shows the day of a recently collected daily assessment with timestamp, location, mood ratings and average heart rate. The watch app shows the flow through the interfaces where the current valence and arousal states can be selected, as well, as the submission screen where participants can opt-out of location logging due to privacy reasons.

Appendix D

EmoRate Study – Participant Material

The following pages contain material used during the *EmoRate* pilot study (Section 5.1). The material comprises:

- Consent form (sample export)
- Ethics approval letter

StressAssess Study Consent

Welcome

This study is aimed to get insights of the relationship between mobile phone/smartwatch sensing data and emotional experiences throughout the day; and it is delivered through an iOS and Apple watch app. The core of this study are emotion self-assessments. You are asked to self-assess your current emotional state by using two commonly used scales: the valence/positivity (negative-sad to positive-happy) of your emotion and the arousal (low-bored/relaxed to high-excited/tense). Furthermore, we collect mobile phone and smartwatch sensing data. This data includes:

• Your location (this is optional and you can easily hide your location from within the app)
• Your Heart Rate for half an hour before the self-assessment
• Your physical activity in form of a step count for half an hour before the self-assessment
• Wrist movements during the self-assessment
• Ambient noise during the self-assessment. This does not include any voice recordings, but just the level of noise in form of a number.

A pre-requisite for this study, is the ownership of an iOS mobile phone and an Apple smartwatch. The study will last 10 days. After you filled the consent form in the app, you are presented with a general questionnaire on demographics and a questionnaire focusing on your perception of stress. Filling out these questionnaires is voluntary.

During the following days, you will get up to 5 notifications a day on your iPhone and Apple Watch at random points during the day. These notifications will ask you to self-assess your current emotional state on your smartwatch. By clicking on the notification on your Apple Watch, you will be automatically guided to the interface. You also have the option to get another reminder in 30 minutes or dismiss and ignore this notification. We highlight, that in the case of any discomfort and unease regarding the logging of your current emotional states, you should not fill out the emotional self-assessment.

This emotion self-assessment takes around 10-30 seconds. You are asked to select a positivity of your current emotion on a scale from very sad to very happy; and you are asked to rate your current arousal level on a scale ranging from very relaxed/bored to very excited/tense. In both cases, you have the option to not give an answer. In the last step of the self-assessment, you see a summary and can also decide if you want to log your current location or not.

During the study you are free to review your past self-assessments in the mobile phone app. If you later wish to delete one of your self-assessments, you can do so in the mobile phone app. After 10 days the study finishes and you are asked to fill out a final questionnaire on your experience with the app and your perceived stress.

Data Gathering

The data for this study is gathered through small assessments on your Apple Watch. These assessments ask you for your current emotional level (e.g. very tense, happy, sad, ...) and will take about 10 seconds to complete. Additional we collect:

• Your location (this is optional and you can easily hide your location from within the app)
• Your Heart Rate for half an hour before the self-assessment
• Your physical activity in form of a step count for half an hour before the self-assessment
• Wrist movements during the self-assessment
• Ambient noise during the self-assessment. This does not include any voice recordings, but just the level of noise in form of a number.

Privacy

The data collected from you is stored on your phone. At the end of the study, we collect the data by connecting your phone to a PC. Your data is anonymized not shared with a third party and can be just accessed by researchers of Queen Mary University London.

Data Use

The data collected from you is used to understand how current emotional states can be picked up using mobile and wearable sensing.

Time Commitment

You will be asked to complete a small assessment on your Apple Watch three to five times a day. Each assessment takes around 10-30 seconds. You can dismiss or ignore the notification, if you do not have time for the self-assessment or do not feel comfortable logging your emotion right now. Additionally, you will be asked to fill out a small survey on perceived stress at the beginning and end of the study.

Withdrawing

You can withdraw from this study at any time without giving any reason.

Katrin Hansel		30/08/2016
Participant's Name (printed)	Participant's Signature	Date

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Hamed Haddadi
CS 404
Department of Computer Science
Queen Mary University of London
Mile End Road LONDON

13th February 2017

To Whom It May Concern:

Re: QMERC2016/17 – Assessment of daily emotional levels with a smartwatch.

The above study was conditionally approved by The Queen Mary Ethics of Research Committee (Panel D) on the 4th May 2016; full approval was ratified via Chair's Action on the 30th June 2016.

This approval is valid for a period of two years, (if the study is not started before this date then the applicant will have to reapply to the Committee).

<u>Amendment</u>
An amendment (small incentive to participate and slightly amended questions) was approved via Chair's Action on the 13th February 2017

Yours faithfully

Ms Elizabeth Hall – QMERC Chair.

Appendix E

Laboratory Experiment – Participant Material

The following pages contain material used during the laboratory experiment (Section 5.2). The material comprises:

- Consent form
- Information sheet
- Ethics approval letter
- Initial questionnaire with demographics, *International Physical Activity Questionnaire (IPAQ)* fitness scale, smoking assessment
- Mood questionnaire with the *Self-Assessment Manikin*, wake/tense arousal and perceived stress assessments
- Provided material on sensor placement

Consent Form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: Assessment of daily emotional levels with a smartwatch
Queen Mary Ethics of Research Committee Ref: QMREF1582

- Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.
- If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.
- I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately.
- By signing this form, I am allowing the researcher to audio or video tape me as part of this research.
- I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.
- I understand, that I will be reimbursed at the end of the experiment with £15 in cash for my time and that I will have to sign for the receipt.

Participant's Statement:

I ____ agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed: Date:

Investigator's Statement:

I ____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer.

Signed: Date:

Participant Statement of Reimbursement:

I ______________________________________ confirm that I received £15 in cash as reimbursement for my time.

Signed: Date:

Pro forma information sheet and consent form

Queen Mary
University of London

Information sheet

Evaluation of Consumer Wearable Devices for Lab-Controlled emotion recognition

We would like to invite you to be part of this research project, if you would like to. You should only agree to take part if you want to, it is entirely up to you. If you choose not to take part there won't be any disadvantages for you and you will hear no more about it.
Please read the following information carefully before you decide to take part; this will tell you why the research is being done and what you will be asked to do if you take part. Please ask if there is anything that is not clear or if you would like more information.
If you decide to take part you will be asked to sign the attached form to say that you agree.
You are still free to withdraw at any time and without giving a reason.

Purpose of the Study

The purpose of the study is to evaluate the consumer wearable fitness devices (Apple Watch, Microsoft Band 2, Polar H7) as a suitable wearable devices to infer emotions during a cognitively demanding task in a controlled lab environment.
The two selected and widely used consumer devices are equipped with promising sensors, such as heart rate and Galvanic Skin Response Sensors. We want to test the reliability of these consumer sensors compared to established, professional-grade sensors (Nexus Kit).

Study Layout

Preparation:

Before we equip you with the sensors, we will ask you to fill out a pre-study questionnaire. This questionnaire contains questions about your demographics (age, gender), experience with wearable technology, smoking behaviour and fitness level (since those are factors which can influence your heart rate readings).
In preparation of the study, you will be equipped with the sensors we use during the experiment. These are 2 wrist-worn wearables (Apple Watch, Microsoft Band 2) which will be placed at your non-dominant arm. A commercial, heart rate chest strap (Polar H7). Furthermore, you will be asked to wear two sensors from the Nexus Kit:

- The Electrocardiogram sensor consists of 3 electrodes: a positive, negative and ground electrode. Those electrodes are placed at the forearm.
- The Galvanic Skin Response Sensor consists of 2 electrodes placed at the index and middle finger of the non-dominant arm.
- A skin temperature sensor will be placed at your upper arm.

All those sensors will be shown and explained to you beforehand.

Study:
The experiment will be conducted in several phases which are repeated. The researcher will be in the room with you throughout the experiment for the instructions and any questions.
Baseline/Relaxed Recording: You will be asked to sit in front of a screen with a slideshow of neutral landscape photos and listen to relaxing music via headphones. This enables us to collect a sensor sample in a relaxed state.
Test Condition: You will be asked to solve arithmetic math equations shown on a screen. You will have to say out loud the answer and the system will provide feedback if the given answer is correct.
The test condition will be performed two times. One time you will be asked to stand in front of the screen. The other time, you will be asked to slowly walk on a treadmill. After each task, you will be asked to shortly fill out a questionnaire about your current mood and how stressful you perceived this task.
At the end of the study, the sensors will be removed.

Audio/Video Recordings

This study involves the audio or video recording throughout the experiment. Neither your name nor any other identifying information will be associated with the audio or audio recording. Only the research team will be able to listen (view) to the recordings. The tapes will be matched with the sensor data by the researcher and erased once the this process is done. Neither your name nor any other identifying information (such as your voice or picture) will be used in presentations or in written products resulting from the study without your prior consent.

Time Commitment

The study will last around 75 - 90 Minutes including introduction, preparation and removal of the sensors in the end.

Reimbursement

For taking part in the experiment, you will be reimbursed with £15 for your time.

Data Sharing

All the information that we collect about you during the course of the study will be kept strictly confidential. You will not be able to be identified or identifiable in any reports or publications. Any data collected about you in the online questionnaire will be stored on a password protected hard drive. Data collected may be shared in an anonymized form to allow reuse by the research team and other third parties. These anonymized data will not allow any individuals or to be identified or identifiable.

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form.
If you wish to contact the responsible research team, please email Katrin Hänsel:
k.hansel@qmul.ac.uk

If you have any questions or concerns about the manner in which the study was conducted please, in the first instance, contact the researcher responsible for the study. If this is unsuccessful, or not appropriate, please contact the Secretary at the Queen Mary Ethics of Research Committee, Room W117, Queen's Building, Mile End Campus, Mile End Road, London or research-ethics@qmul.ac.uk.

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Hamed Haddadi
Room CS 404
Department of Computer Science
Queen Mary University of London
Mile End Road
London

22nd May 2017

To Whom It May Concern:

Re: QMREC1582 – Evaluation of Consumer Wearables for Emotion Recognition in a Controlled Lab Environment.

I can confirm that Katrin Hänsel has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that her proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Ms Hazel Covill – QMERC Administrator

Date/Time: ____________
Participant ID: ____________

Demographic Data and Assessment of Influential Factors

Age: ____

Gender: O male
O female
O non-binary
O I prefer not to tell

Do you own/use a (fitness) wearable device e.g. Microsoft Band, Apple Smartwatch, Fitbit, Jawbone, etc.? If yes, please tell us which.

____

IPAQ - INTERNATIONAL PHYSICAL ACTIVITY QUESTIONNAIRE (Short)

We are interested in finding out about the kinds of physical activities that people do as part of their everyday lives. The questions will ask you about the time you spent being physically active in the last 7 days. Please answer each question even if you do not consider yourself to be an active person. Please think about the activities you do at work, as part of your house and yard work, to get from place to place, and in your spare time for recreation, exercise or sport.
Think about all the vigorous activities that you did in the last 7 days. Vigorous physical activities refer to activities that take hard physical effort and make you breathe much harder than normal. Think only about those physical activities that you did for at least 10 minutes at a time.

1. During the last 7 days, on how many days did you do vigorous physical activities like heavy lifting, digging, aerobics, or fast bicycling?
_____ days per week
No vigorous physical activities Skip to question 3

2. How much time did you usually spend doing vigorous physical activities on one of those days?
_____ hours per day
_____ minutes per day
Don’t know/Not sure

Date/Time: ____________
Participant ID: ____________

Think about all the moderate activities that you did in the last 7 days. Moderate activities refer to activities that take moderate physical effort and make you breathe somewhat harder than normal. Think only about those physical activities that you did for at least 10 minutes at a time.

3. During the last 7 days, on how many days did you do moderate physical activities like carrying light loads, bicycling at a regular pace, or doubles tennis? Do not include walking.
_____ days per week
No moderate physical activities Skip to question 5

4. How much time did you usually spend doing moderate physical activities on one of those days?
_____ hours per day
_____ minutes per day
Don't know/Not sure

Think about the time you spent walking in the last 7 days. This includes at work and at home, walking to travel from place to place, and any other walking that you have done solely for recreation, sport, exercise, or leisure.

5. During the last 7 days, on how many days did you walk for at least 10 minutes at a time?
_____ days per week
No walking Skip to question 7

6. How much time did you usually spend walking on one of those days?
_____ hours per day
_____ minutes per day
Don't know/Not sure

The last question is about the time you spent sitting on weekdays during the last 7 days. Include time spent at work, at home, while doing course work and during leisure time. This may include time spent sitting at a desk, visiting friends, reading, or sitting or lying down to watch television.

7. During the last 7 days, how much time did you spend sitting on a weekday?
_____ hours per day
_____ minutes per day
Don't know/Not sure

Date/Time: ____________
Participant ID: ____________

What kind of sports do you perform regularly?

____

SMOKING BEHAVIOR

Please circle the answer most relevant to you

What is your current cigarette smoking behavior (including hand-rolled cigarettes)?

A Daily smoker (at least one cigarette per day, disregarding religious fasting)
B Occasional smoker (less than one cigarette per day)
C Ex-smoker of cigarettes
D Non-smoker of cigarettes
O I prefer not to tell

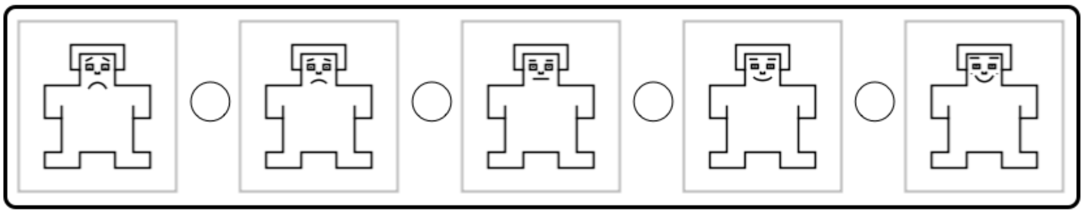
This is the end of the questionnaire, thank you for participating!

Date/Time: ____________
Participant ID: ____________
Experiment Stage: ____________

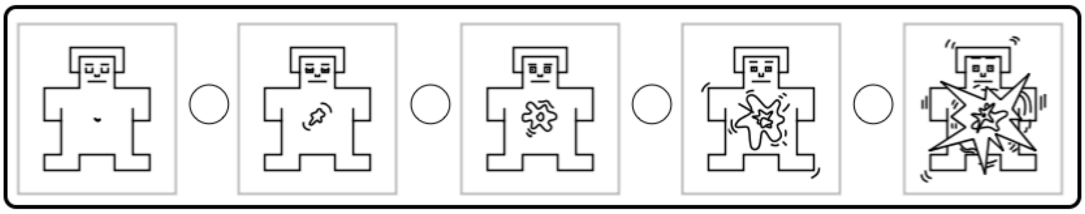
Mood Questionnaire

Please select one option each to describe your current mood:

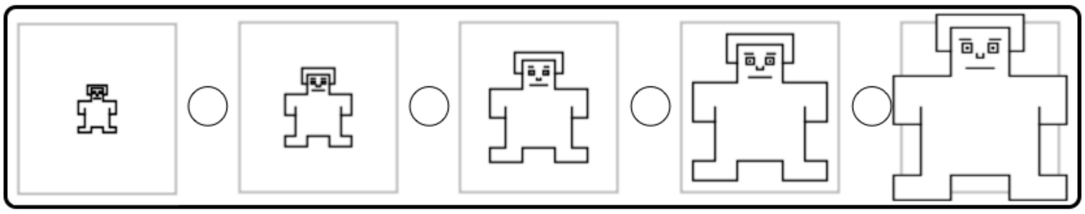
Valence (Unhappy vs Happy)



Arousal (Calm vs Excited)



Dominance (Controlled vs In-Control)



Please circle the answer to these two questions:

How wide-awake do you feel?

○ ○ ○ ○ ○ ○

not at all | very much

How nervous do you feel?

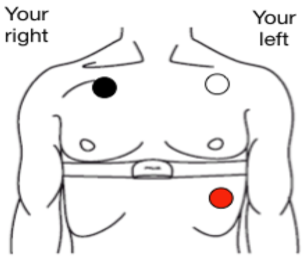
○ ○ ○ ○ ○ ○

not at all | very much

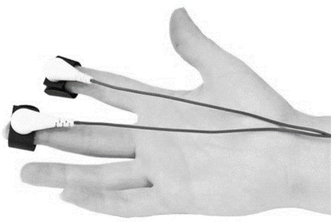
How stressful did you perceive the previous task?

○ ○ ○ ○ ○ ○

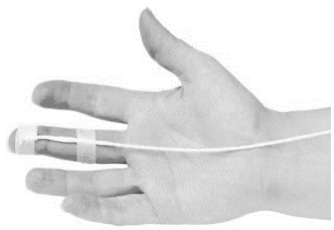
not at all | very much



(a) Placement instructions for the H7 Polar chest belt and three pre-gelled, disposable electrodes from the Nexus-10.



(b) Placement instructions for the two skin conductance recording electrodes from the Nexus-10.



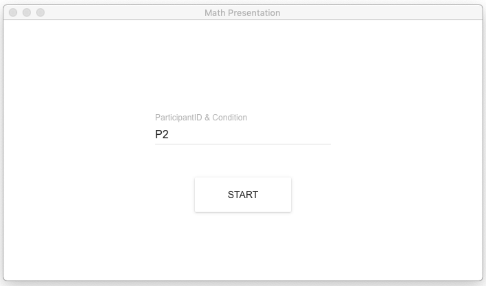
(c) Placement instructions for the skin temperature sensor from the Nexus-10.

Figure A.10.: These three figures were presented to the participants helping them to correctly place the sensors, namely the H7 Polar chest belt and three pre-gelled, disposable electrodes for the ECG signal, two skin conductance recording electrodes, and another skin temperature sensor; all of these belonging to the Nexus-10 (Figure (a) was taken and modified from the Polar manufacturer website; Figures (b) and (c) are taken from Nexus-10 manufacturer's website[6]).

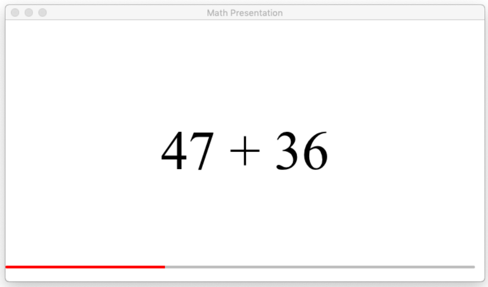
Appendix F

Laboratory Experiment – MAT Experimenter and Participant UI

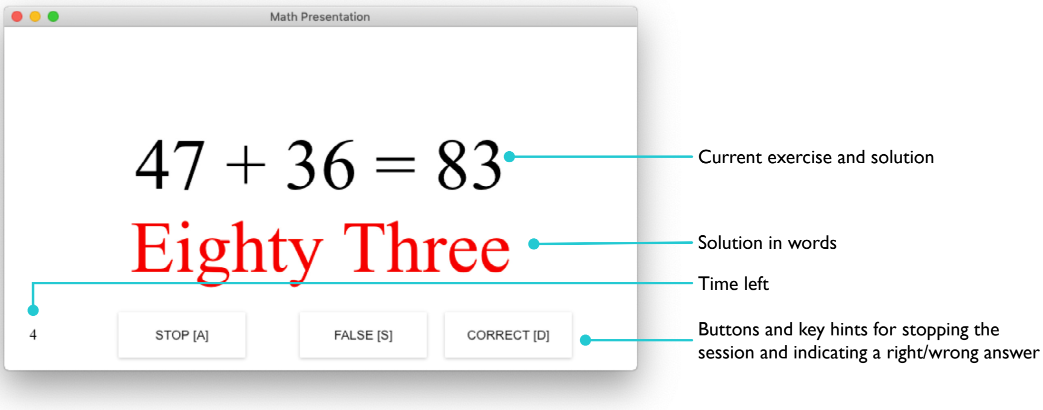
The below application was used to present the *Mental Arithmetic Tasks (MAT)* stimulus to participants in Section 5.2. Source code available: `https://github.com/MiezelKat/LabMAT`.



(a) Start Screen.



(b) Participant view.



(c) Experimenter view

Figure A.11.: Screenshots of the *MAT* application during the laboratory experiment

Appendix G

Laboratory Experiment – Summary of Data

The following table contains the descriptive statistics of all the collected sensory and subjective measurements of the laboratory experiment from Section 5.2.

		Relaxed				MAT			
		Stationary		Treadmill		Stationary		Treadmill	
		Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Nexus	Heart Rate	66.16 (10.6)	63.73	85.55 (10.48)	89.26	69.5 (10.1)	68.81	88.93 (10.15)	89.13
	Skin Conductance	13.38 (50.26)	1.91	14.99 (49.94)	3.35	6.21 (10.51)	3.9	15.06 (49.92)	3.44
	Skin Temperature	32.1 (2.64)	32.16	30.76 (1.85)	30.98	31.14 (2.04)	31.58	30.63 (1.85)	30.97
Apple Watch	Heart Rate	66.27 (10.64)	64.04	94.86 (19.17)	93.45	68.83 (9.93)	68.41	98.61 (18.97)	95.25
Microsoft Band	Heart Rate	66.55 (9.66)	64.82	72.38 (7.62)	73.56	68.44 (7.62)	68.6	72.82 (8.75)	73.48
	Skin Conductance	0.32 (0.38)	0.19	0.74 (1.05)	0.43	0.8 (1.14)	0.48	0.79 (1.13)	0.39
	Skin Temperature	30.02 (1.51)	30.47	29.63 (1.77)	29.75	29.97 (1.88)	29.87	29.34 (1.55)	29.12
Polar	Heart Rate	66.89 (11.26)	66.24	87.04 (11.41)	88.08	69.06 (10.87)	69	89.3 (13.38)	88.8
Subjective Measures	Arousal	-2.76 (1.14)	-3	-1.76 (1.64)	-2	0 (2.28)	0	0.81 (1.99)	1
	Wake Arousal	2.29 (1.06)	2	3.05 (0.97)	3	3.52 (0.93)	3	3.86 (0.79)	4
	Tense Arousal	0.71 (1.01)	0	0.86 (0.73)	1	1.52 (1.17)	1	1.71 (1.19)	2
	Valence	1.19 (1.47)	1	0.9 (1.14)	1	0.95 (1.69)	1	1.19 (1.63)	2
	Dominance	0.48 (1.6)	0	0.67 (1.32)	0	-0.38 (2.31)	0	0 (2.32)	0
	Perceived Stress	0.24 (0.54)	0	0.86 (0.96)	1	2.33 (1.39)	3	2.57 (1.4)	3

Table A.1.: Overview of the descriptive statistics of sensor and subjective measures of the laboratory experiment.

Appendix H

Design Space Evaluation - Participant Material

The following pages contain material used during the evaluation of the *Design Space for Physiological Measurement Tools* (Chapter 6). The material comprises:

- Consent form
- Ethics approval letter
- Overview of the initial five dimensions of the design space as it has been shown to participants during the interviews (Figure A.12)

University of Stuttgart
Germany

Queen Mary
University of London

Consent Form

DESCRIPTION: You are invited to participate in **an interview aiming to explore your personal experiences with sensing technologies**.

TIME INVOLVEMENT: Your participation will take approximately **30 minutes.**

DATA COLLECTION: For this interview you will be asked to state your opinions and share your personal experiences in working with physiological sensing and sensing technologies. Furthermore, we will record this interview session to be able to listen to your statements again.

RISKS AND BENEFITS: No risk is associated with this interview. The collected data is securely stored. We do guarantee no data misuse and privacy is completely preserved. Your data will be used anonymously and for scientific purpose only.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand your **participation is voluntary** and you have the **right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled**. **The alternative is not to participate.** You have the right to refuse to answer particular questions. The results of this research project may be presented at scientific or professional meetings or published in scientific journals. Your identity is not disclosed unless we directly inform and ask for your permission.

CONTACT INFORMATION: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact following persons:

Romina Poguntke (romina.poguntke@vis.uni-stuttgart.de)
Katrin Hänsel (k.hansel@qmul.ac.uk)

By signing this document I confirm that I understand and agree to the terms and conditions.

Name: ____________________ *Signature, Date:* ____________________

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Akram Alomainy
School of Electronic Engineering
and Computer Science
Eng E202
Queen Mary University of London
Mile End
London

29th May 2018

To Whom It May Concern:

Re: QMREC1870 - Qualitative Study on User and Expert Experiences with Wearable Sensing Devices.

I can confirm that Katrin Hänsel has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that her proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Mr Jack Biddle – Research Approvals Advisor

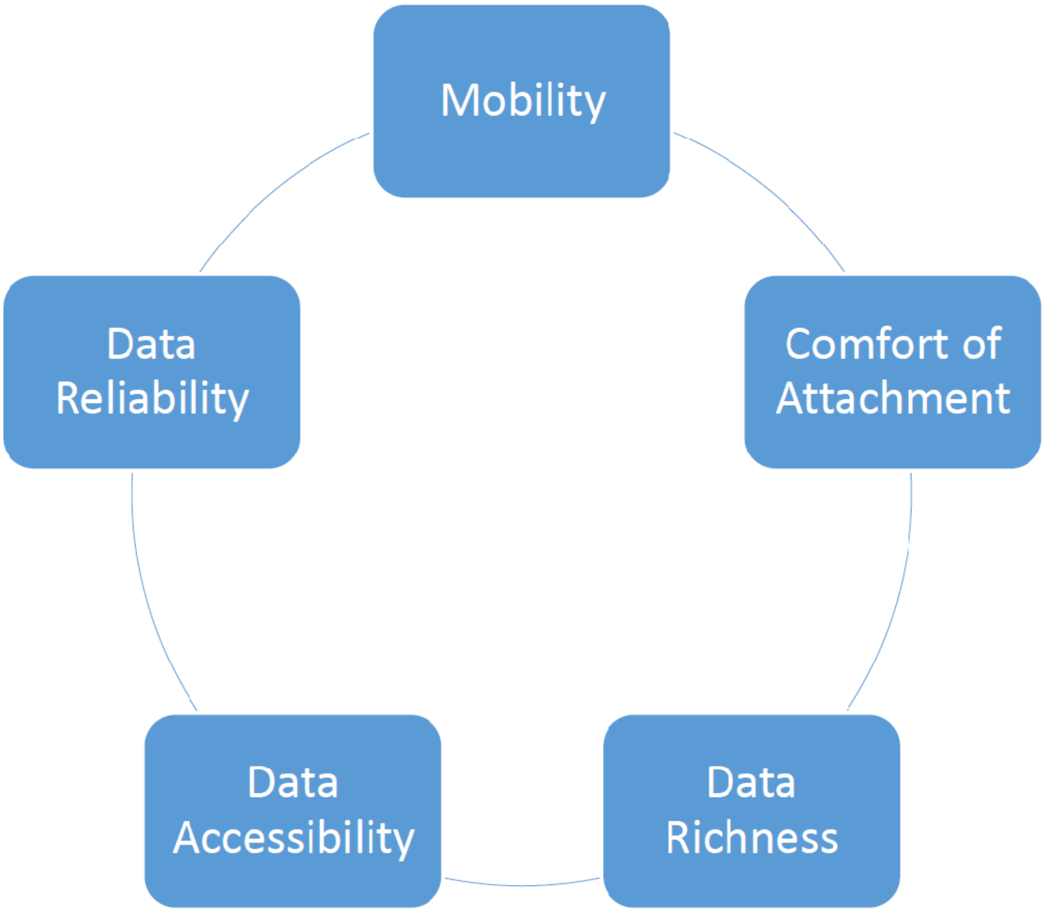


Figure A.12.: Five dimensions of the original design space which was presented to the interviewees

Appendix I

Design Space Evaluation - Participant Demographics

The following two tables show the demographic information of the interviewees from the design space evaluation (Chapter 6).

						Experience with Wearable Sensors		
ID	**Gender**	**Age**	**Nationality**	**Occupation**	**Field**	**years**	**sensors**	**pub.†**
E1	male	27	Macedonian	young researcher + PhD student	affective computing	4 years	Empatica E3 + E4 , Microsoft Band, chest straps	19
E2	female	30	Egyptian	researcher + PhD student	affective computing	4 years	Fitbit Charge 2, Microsoft Band, MoviSense, Muse, Emotiv Epoc, OpenBCI, NeuroSky Mindband	9
E3	female	30	Swiss	young researcher + PhD student	interruptibility	4.5 years	Empatica E3 + E4, Affectiva Q Sensor, Polar H7, Fitbit Charge 2	7
E4	male	34	German	researcher + PhD student	biomedical engineering	4 years	Empatica E3, Polar H7, Shimmer sensors, Sony SmartBand, Nexus Kit-10	7
E5	male	32	Liechtenstein	software developer (former PhD student & post-doc researcher)	software engineering	4 years	Empatica E3 + E4, Affectiva Q Sensor, Fitbit Charge 2, Polar H7, SenseCore chest band, NeuroSky Mindband	14

Table A.2.: Demographic information of the experts interviewed; † pub. abbrev. for number of publications

Appendix J

Interaction Study - Participant Material

The following pages contain material used during the laboratory experiment (Chapter 7). The material comprises:

- Consent form
- Ethics approval letter
- Initial questionnaire with demographics,
- Mood questionnaire with the *Self-Assessment Manikin*, *Positive Negative Affect Schedule (PANAS)*
- Initial relationship questionnaire with *Inclusion of Others in Self (IOS) Scale*
- Interaction questionnaire administered mid-study

questionnaire name	time administered	type/ used scales	collected data
Initial Questionnaire	- pre-study (5 days prior)	demographics	age, gender, employment
		EQ Scale	empathy score (8 to 32)
		BFI-10	score for openness, conscientiousness, extraversion, neuroticism, agreeableness (2 to 10 each)
Relationship Questionnaire	- pre-experiment - post-experiment (12 h after)	*IOS Scale*	bi-directed score for closeness (1 to 7)
		sympathy scale	bi-directed sympathy score (1 to 10)
		relationship	bi-directed self-described relationship
Mood Questionnaire	- pre-experiment (after baseline) - mid-experiment - post-experiment	*PANAS*	score for positive and negative affect (10 to 50 each)
		SAM	score for valence, arousal, dominance (-4 to 4)
Interaction Questionnaire	- mid-experiment - post-experiment (12 h after)	interaction quality	bi-directed ordinal score for quality of interaction (good, neutral, bad or no interaction)

Table A.3.: Overview of the utilised questionnaires and scales used during the study

Consent Form

Mobile and Wearable Sensing in a Social Mingling Scenario

Queen Mary Ethics of Research Committee Ref: [TODO]

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

- Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.
- If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.
- I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately.
- By signing this form, I am allowing the researcher to audio or video tape me as part of this research.
- I consent to the processing of my personal information for the purposes of this research study. This data will be shared with researchers involved in this project and an anonymised version may be made available for other researchers and academic purposes. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.
- I understand, that I will be reimbursed at the end of the experiment with £20 for my time and that I will have to sign for the receipt.

Participant's Statement:

I ____ agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed: Date:

Investigator's Statement:

I ____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer.

Signed: Date:

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Akram Alomainy
Eng E202, Engineering building
Electronic Engineering & Computer Science
Mile End
London

22nd November 2017

To Whom It May Concern:

Re: QMREC1705 - Mobile and Wearable Sensing in a Social Mingling Scenario.

I can confirm that Katrin Hansel has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that her proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Mr Jack Biddle – Research Approvals Advisor

Mood Questionnaire

This scale consists of a number of words that describe different feelings and emotions. Read each item and then list the number from the scale below next to each word. Next to each item, please indicate to what extent you feel this way right now, that is, at the **present moment**.

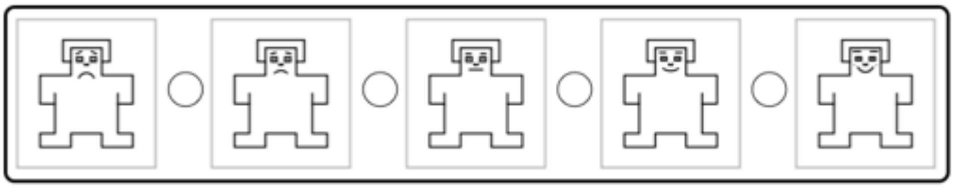
1	2	3	4	5
Very slightly or not at all	A little	Moderately	Quite a bit	Extremely

Interested	Irritable
Distressed	Alert
Excited	Ashamed
Upset	Inspired
Strong	Nervous
Guilty	Determined
Scared	Attentive
Hostile	Jittery
Enthusiastic	Active
Proud	Afraid

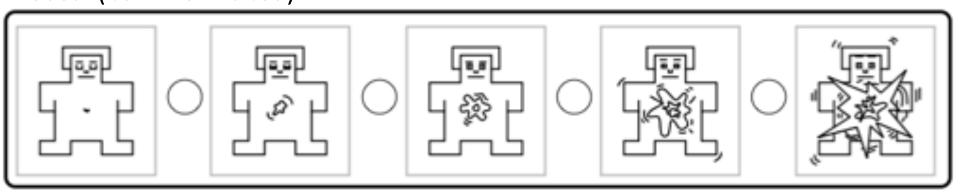
Self Assessment Manikin

Please use the following scales to describe your current mood. For each dimension (valence, arousal, dominance) indicate the level how you feel. The small prints under each scale are examples of emotions related to that side of the scale. Place you X either on a Manikin or if you feel between two Manikins, place the X in the circle.

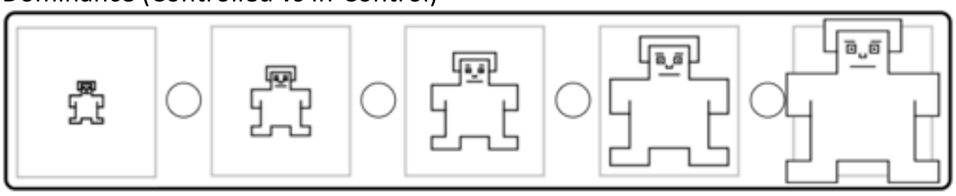
Valence (Unhappy vs Happy)



Arousal (Calm vs Excited)



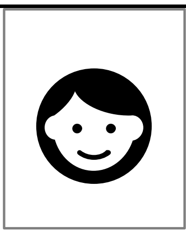
Dominance (Controlled vs In-Control)

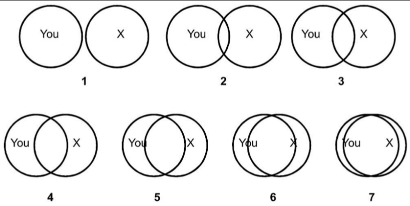


Relationship Questionnaire after Study

Below, you will find a table with all the participants in the experiment. Please fill out the short questionnaire for each of them indicating your relationship with them and your feeling of **sympathy/antipathy**. Please tick the boxed which match the closest.
In the column 'closeness', you find a scale for how **close you feel to the participant X**. Please pick one of the 7 options matching your perception of closeness to participant X. Please, indicate if you have interacted. If you have interacted, indicate the **pleasantness of interaction**. Don't fill out next to your own photo.

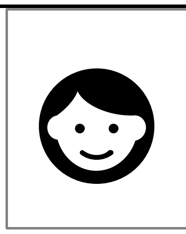
P2

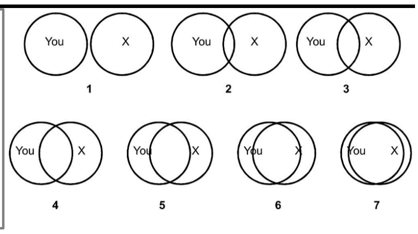




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

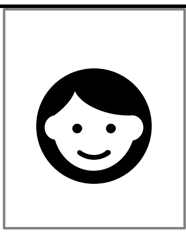
P6

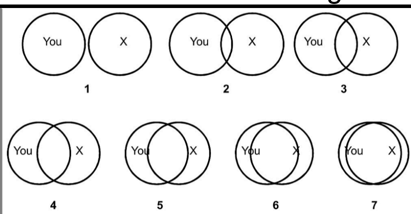




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

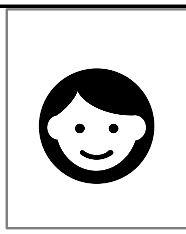
P4

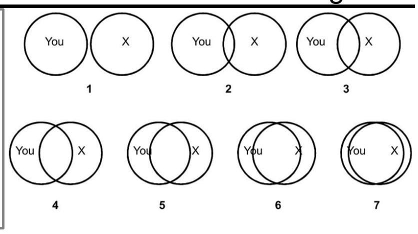




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

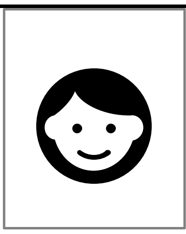
P8

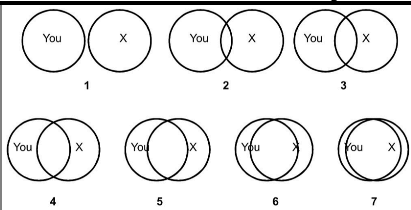




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

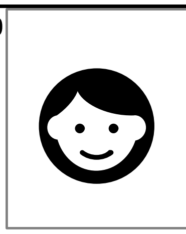
P5

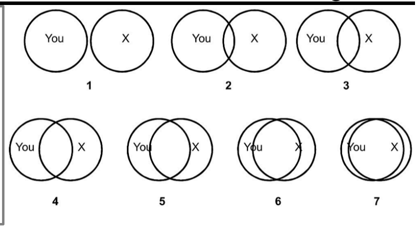




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

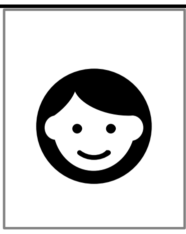
P10

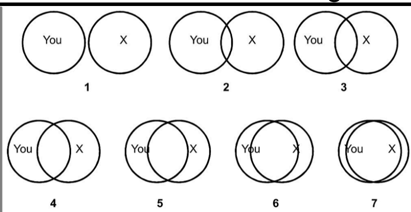




Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

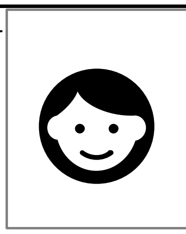
P12

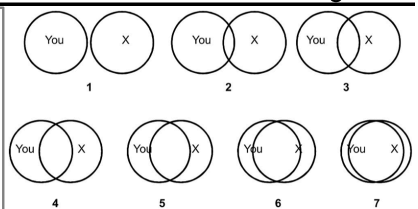




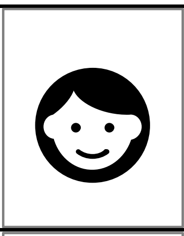
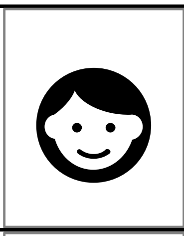
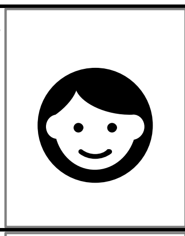
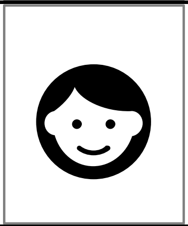
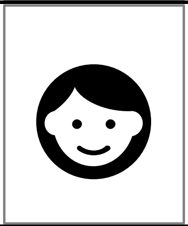
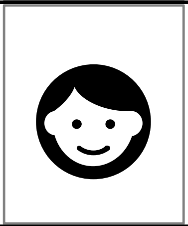
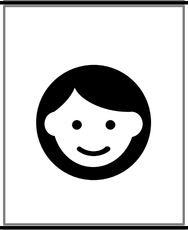
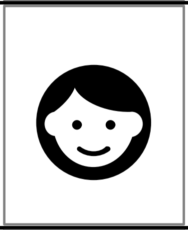
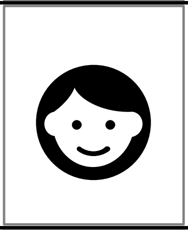
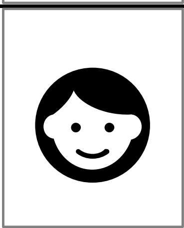
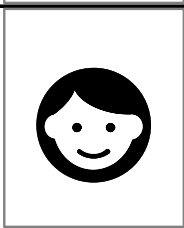
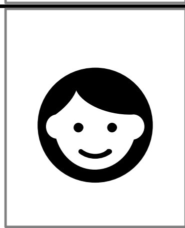
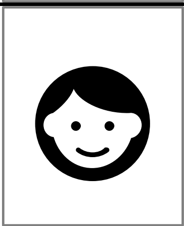
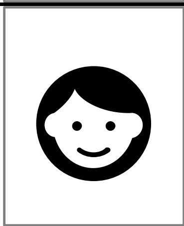
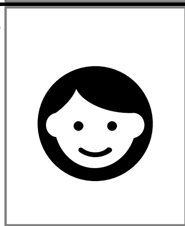
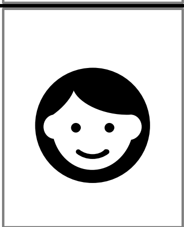
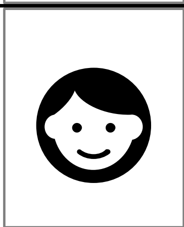
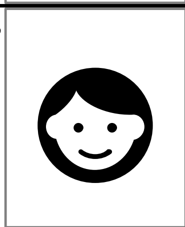
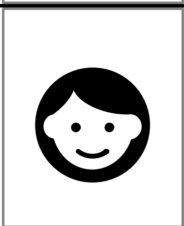
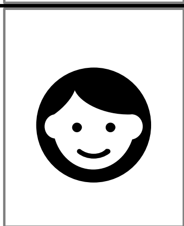
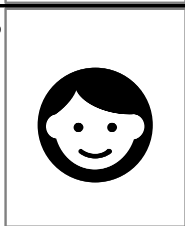
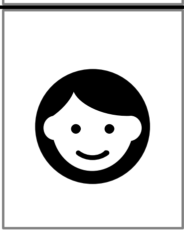
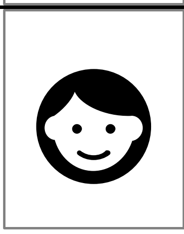
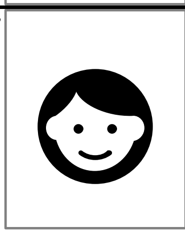
Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

P14





Sympathy: low o o o o o o o o o o high
Relationship: ________________
Interacted: o yes o no
Pleasantness: low o o o o o o o o o o high

P2		Interaction: o good o neutral o bad o none	P6		Interaction: o good o neutral o bad o none	P33		Interaction: o good o neutral o bad o none
P4		Interaction: o good o neutral o bad o none	P8		Interaction: o good o neutral o bad o none	P37		Interaction: o good o neutral o bad o none
P5		Interaction: o good o neutral o bad o none	P10		Interaction: o good o neutral o bad o none	P39		Interaction: o good o neutral o bad o none
P12		Interaction: o good o neutral o bad o none	P14		Interaction: o good o neutral o bad o none	P9		Interaction: o good o neutral o bad o none
P19		Interaction: o good o neutral o bad o none	P22		Interaction: o good o neutral o bad o none	P35		Interaction: o good o neutral o bad o none
P23		Interaction: o good o neutral o bad o none	P24		Interaction: o good o neutral o bad o none	P38		Interaction: o good o neutral o bad o none
P27		Interaction: o good o neutral o bad o none	P29		Interaction: o good o neutral o bad o none	P26		Interaction: o good o neutral o bad o none
P30		Interaction: o good o neutral o bad o none	P31		Interaction: o good o neutral o bad o none	P17		Interaction: o good o neutral o bad o none

Appendix K

Interaction Study – Floorplan

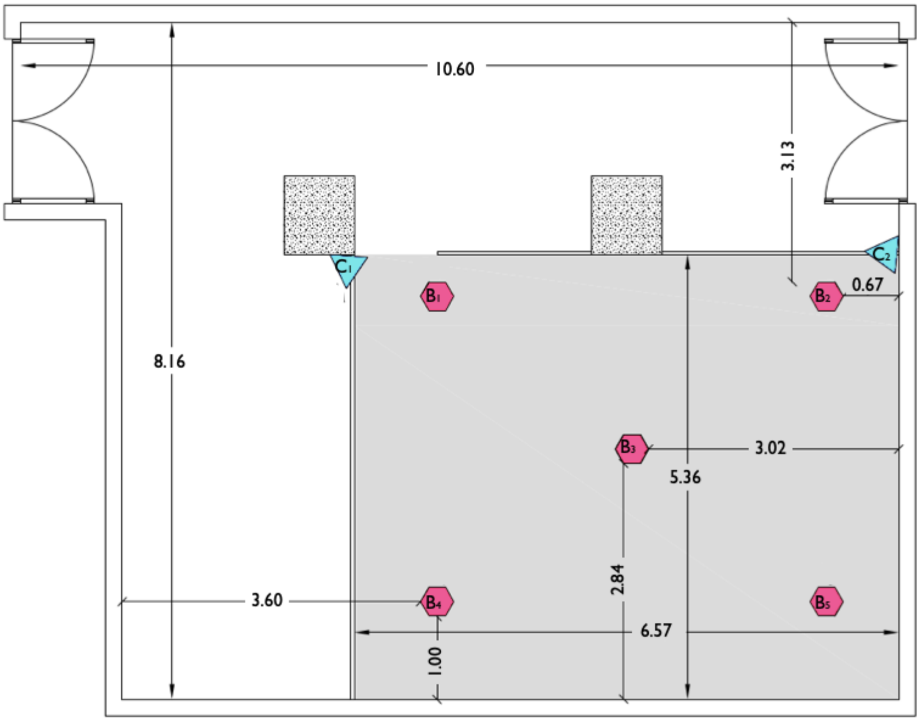


Figure A.13.: Floor plan of the Performance Lab. Marked are the placement of Bluetooth Beacons (B) at the ceiling and the cameras (C_1 and C_2) used for ground truth recording. (distances in meters)

Bibliography

Ackoff, R.L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, **16**:3–9.

Ahn, J.M. (2013). Heart rate variability (hrv) analysis using simultaneous handgrip electrocardiogram and fingertip photoplethysmogram. *Advances in Information Sciences and Service Sciences*, **5**(13):164–170. doi:10.1117/12.2045530.

Aiello, L.M., Quercia, D., and Roitmann, E. (2018). Hearts and politics: Metrics for tracking biorhythm changes during brexit and trump. In *Proceedings of the 2018 International Conference on Digital Health*, DH '18, pp. 111–115. ACM, New York, NY, USA. doi:10.1145/3194658.3194678.

Akyazi, O., Batmaz, S., Kosucu, B., and Arnrich, B. (2017). Smokewatch: A smartwatch smoking cessation assistant. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. doi:10.1109/SIU.2017.7960536.

Alberto, F.P., Nathanael, M., Mathew, B., and Ainsworth, B.E. (2017). Wearable monitors criterion validity for energy expenditure in sedentary and light activities. *Journal of Sport and Health Science*, **6**(1):103 – 110. doi:10.1016/j.jshs.2016.10.005.

Alcorta, C.S. and Sosis, R. (2005). Ritual, emotion, and sacred symbols. *Human Nature*, **16**(4):323–359. doi:10.1007/s12110-005-1014-3.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, **28**(3):R1–R39. doi:10.1088/0967-3334/28/3/R01.

Althoff, T., Sosič, R., Hicks, J.L., King, A.C., Delp, S.L., and Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, **547**(7663):336–339. doi:10.1038/nature23018.

Altini, M. (2013). Heart rate variability logger - app details. accessed 07/05/2019.
URL https://www.marcoaltini.com/blog/heart-rate-variability-logger-app-details

Andronico, M. (2017). Microsoft band 2 vs. apple watch, fitbit surge and garmin vivoactive. accessed 15/12/2017.
URL https://www.tomsguide.com/us/microsoft-band-vs-apple-watch,news-21684.html

Apple (2015). Researchkit. accessed 09/10/2015.
URL https://www.apple.com/researchkit/

Apple (2016). Your heart rate. what it means, and where on apple watch you'll find it. accessed 25/11/2016.
URL https://support.apple.com/en-gb/HT204666

Apple Inc. (2017). Apple watch series 3 battery information. accessed 15/12/2017.
URL https://www.apple.com/uk/watch/battery.html

Apple Inc. (2018). Developer documentation - core motion. accessed 17/10/2018.
URL https://developer.apple.com/documentation/coremotion

Apple Inc. (2019a). Ecg app and irregular heart rhythm notification available today on apple watch. accessed 07/05/2019.
URL https://www.apple.com/uk/newsroom/2019/03/ecg-app-and-irregular-rhythm-notification-on-apple-watch-available-today-across-europe-and-hong-kong/

Apple Inc. (2019b). ios - health - apple (uk). accessed 07/05/2019.
URL https://www.apple.com/uk/ios/health/

Aron, A., Aron, E.N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, **63**(4):596–612. doi:10.1037//0022-3514.63.4.596.

de Arriba-Pérez, F., Caeiro-Rodríguez, M., and Santos-Gago, J.M. (2016). Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors*, **16**(9). doi:10.3390/s16091538.

Ball, M. (2015). Qs access: Personal data freedom. accessed 10/02/2019.
URL http://quantifiedself.com/data-ownership/

Bandodkar, A.J., Jeerapan, I., and Wang, J. (2016). Wearable chemical sensors: Present challenges and future prospects. *ACS Sensors*, **1**(5):464–482. doi:10.1021/acssensors.6b00250.

Baron-Cohen, S. and Wheelwright, S. (2004). The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, **34**:163 – 175. doi:10.1023/B:JADD.0000022067.19833.0.

Ben Abdesslem, F., Phillips, A., and Henderson, T. (2009). Less is more: Energy-efficient mobile sensing with senseless. In *Proceedings of the 1st ACM Workshop on Networking, Systems, and Applications for Mobile Handhelds*, MobiHeld '09, pp. 61–62. ACM, New York, NY, USA. doi:10.1145/1592606.1592621.

Ben-Shakhar, G. (1985). Standardization within individuals: a simple method to neutralize individual differences in skin conductance. *Psychophysiology*, **22**(3):292–299. doi:10.1111/j.1469-8986.1985.tb01603.x.

Bernard, H.R. (2006). *Research Methods in Anthropology, 4th edition*. AltaMira Press.

Bernieri, F.J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, **12**(2):120–138. doi:10.1007/BF00986930.

Biederman, I. (1973). Mental set and mental arithmetic. *Memory & Cognition*, **1**(3):383–386.

Bluetooth Special Interest Group (2010). Specification of the bluetooth system, covered core package, version 4.0.

Bluetooth Special Interest Group (2011). Heart rate profile specification. accessed 07/05/2019.
URL https://www.bluetooth.org/docman/handlers/downloaddoc.ashx?doc_id=239865

Bluetooth Special Interest Group (2018). Gatt overview. accessed 17/10/2018.
URL https://www.bluetooth.com/specifications/gatt/generic-attributes-overview

Blythe, A., Crouter, S.E., and LaMunion, S.R. (2017). *Validity of Consumer-Based Physical Activity Monitors for Estimating Energy Expenditure in Youth*. Master's thesis, University of Tennessee, Knoxville, USA.

Bodine, K. and Gemperle, F. (2003). Effects of functionality on perceived comfort of wearables. *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pp. 57–61. doi:10.1109/ISWC.2003.1241394.

Boettger, S., Puta, C., Yeragani, V.K., Donath, L., Müller, H.J., Gabriel, H.H.W., and Bär, K.J. (2010). Heart rate variability, qt variability, and electrodermal activity during exercise. *Medicine and science in sports and exercise*, **42 3**:443–8.

Boker, S.M., Xu, M., Rotondo, J.L., and King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, **7**(3):338–55.

Bradley, M.M. and Lang, P.J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal Of Behavior Therapy And Experimental Psychiatry*, **25**(1):49–59.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, **3**(2):77–101. doi:10.1191/1478088706qp063oa.

British Cardiovascular Society (2010). Recording a standard 12-lead electrocardiogram — an approved methodology.
URL https://www.bcs.com/documents/consensus_guidelines.pdf

Buechley, L., Eisenberg, M., Catchen, J., and Crockett, A. (2008). The lilypad arduino: Using computational textiles to investigate engagement, aesthetics, and diversity in computer science education. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 423–432. ACM, New York, NY, USA. doi:10.1145/1357054.1357123.

Caldwell, S. and Filipowicz, L. (2018). How to pair an external heart rate monitor to apple watch. accessed 08/05/2019.
URL https://www.imore.com/how-pair-external-heart-rate-monitor-apple-watch

Callister, R., Suwarno, N.O., and Seals, D.R. (1992). Sympathetic activity is influenced by task difficulty and stress perception during mental challenge in humans. *The Journal of Physiology*, **454**(1):373–387. doi:10.1113/jphysiol.1992.sp019269.

Canhoto, A.I. and Arp, S. (2017). Exploring the factors that support adoption and sustained use of health and fitness wearables. *Journal of Marketing Management*, **33**(1-2):32–60. doi:10.1080/0267257X.2016.1234505.

Cappella, J. (1997). Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators. *Departmental Papers (ASC)*, **72**. doi:10.1037/0022-3514.72.1.119.

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.F., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLOS ONE*, **5**(7):1–9. doi:10.1371/journal.pone.0011596.

Chamberlain, A., Crabtree, A., Rodden, T., Jones, M., and Rogers, Y. (2012). Research in the wild: Understanding 'in the wild' approaches to design and development. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pp. 795–796. ACM, New York, NY, USA. doi:10.1145/2317956.2318078.

de Chazal, P., O'Dwyer, M., and Reilly, R.B. (2004). Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, **51**(7):1196–1206. doi:10.1109/TBME.2004.827359.

Chevalier, G. (2016). Lstms for human activity recognition. accessed 08/05/2019.
URL https://github.com/guillaume-chevalier/LSTM-Human-Activity-Recognition

Choe, E.K., Lee, N.B., Lee, B., Pratt, W., and Kientz, J.A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 1143–1152. ACM, New York, NY, USA. doi:10.1145/2556288.2557372.

Clarkson, B., Mase, K., and Pentland, A. (2000). Recognizing user context via wearable sensors. In *Digest of Papers. Fourth International Symposium on Wearable Computers*, pp. 69–75. doi:10.1109/ISWC.2000.888467.

Cohen, S. and Pressmann, S.D. (2006). Positive affect and health. *Current Directions in Psychological Science*, **15**(3).

Collins, F.S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, **372**(9):793–795. doi:10.1056/NEJMp1500523.

Consolvo, S., McDonald, D.W., Toscos, T., Chen, M.Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., and Landay, J.A. (2008). Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1797–1806. ACM. doi:10.1145/1357054.1357335.

Craig, C.L., Marshall, A.L., Sjorstrom, M., Bauman, A.E., Booth, M.L., Ainsworth, B.E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J.F., and Oja, P. (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine and Science in Sports and Exercise*, **35**(8):1381–1395. doi:10.1249/01.MSS.0000078924.61453.

Dameff, C., Clay, B., and Longhurst, C.A. (2019). Personal Health Records: More Promising in the Smartphone Era?Personal Health Records—More Promising in the Smartphone Era?Personal Health Records—More Promising in the Smartphone Era? *JAMA*, **321**(4):339–340. doi:10.1001/jama.2018.20434.

Dargazany, A.R., Stegagno, P., and Mankodiya, K. (2018). Wearabledl: Wearable internet-of-things and deep learning for big data analytics—concept, literature, and future. *Mobile Information Systems*, **2018**. doi:10.1155/2018/8125126.

Davidson, R.J., Jackson, D.C., and Larson, C.L. (2000). Human electroencephalography. *Handbook of psychophysiology*, **2**:27–52.

Dawson, M.E., Schell, A.M., and Filion, D.L. (2007). The electrodermal system. *Handbook of psychophysiology*, **2**:200–223.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, **3**(3):349–365. doi:10.1109/T-AFFC.2012.12.

Di Lascio, E., Gashi, S., and Santini, S. (2018). Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, **2**(3):103:1–103:21. doi:10.1145/3264913.

Dimitroff, S.J., Kardan, O., Necka, E.A., Decety, J., Berman, M.G., and Norman, G.J. (2017). Physiological dynamics of stress contagion. *Scientific Reports*, **7**(6168). doi:10.1038/s41598-017-05811-1.

Dominick, G.M., Winfree, K.N., Pohlig, R.T., and Papas, M.A. (2016). Physical activity assessment between consumer- and research-grade accelerometers: A comparative study in free-living conditions. *JMIR mHealth and uHealth*, **4**(3). doi:10.2196/mhealth.6281.

Dooley, E.E., Golaszewski, N.M., and Bartholomew, J.B. (2017). Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices. doi:10.2196/mhealth.7043.

Doyle, B. (2016). 5 reasons why google glass was a miserable failure read more at https://www.business2community.com/tech-gadgets/5-reasons-google-glass-miserable-failure-01462398. accessed 04/02/2019.
URL `https://www.business2community.com/tech-gadgets/5-reasons-google-glass-miserable-failure-01462398`

Dunn, J., Runge, R., and Snyder, M. (2018). Wearables and the medical revolution. *Personalized Medicine*, **15**(5):429–448. doi:10.2217/pme-2018-0044.

Dunne, L. (2004). The design of wearable technology: Addressing the human-device interface through functional apparel design.

Dunne, L.E., Profita, H., Zeagler, C., Clawson, J., Gilliland, S., Do, E.Y., and Budd, J. (2014). The social comfort of wearable technology and gestural interaction. In *2014 36th Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4159–4162. doi:10.1109/EMBC.2014.6944540.

El-Amrawy, F. and Nounou, M.I. (2015). Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? *Healthcare Informatics Research*, **21**(4):315. doi:10.4258/hir.2015.21.4.315.

Elfenbein, H.A. (2014). The many faces of emotional contagion: An affective process theory of affective linkage. *Organizational Psychology Review*, **4**(4):326–362. doi:10.1177/2041386614542889.

Empatica Inc. (n.d.). E4 wristband.
URL `https://www.empatica.com/en-eu/research/e4/`

Empatica Inc. (2014). Empatica e4 - revision 001.
URL `http://box.empatica.com/documentation/20141119_E4_TechSpecs.pdf`

Empatica Support (2017a). Have you done comparative studies or validation of the heart rate obtained from the e3/e4? accessed 10/06/2019.
URL `https://support.empatica.com/hc/en-us/articles/200293658`

Empatica Support (2017b). Have you done comparative studies or validation on electrodermal activity sensor? accessed 10/06/2019.
URL `https://support.empatica.com/hc/en-us/articles/203005295`

Ericsson Consumerlab (2015). Living longer: Wellness and the internet. Technical report.
URL `http://www.ericsson.com/res/docs/2015/consumerlab/consumerlab-living-longer-wellness-and-the-internet.pdf`

Evenson, K.R., Goto, M.M., and Furberg, R.D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*, **12**(1):159. doi:10.1186/s12966-015-0314-1.

Exler, A., Schankin, A., Klebsattel, C., and Beigl, M. (2016). A wearable system for mood assessment considering smartphone features and data from mobile ecgs. In *Proceedings of* *the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pp. 1153–1161. ACM, New York, NY, USA. doi:10.1145/2968219.2968302.

Feldman, R., Greenbaum, C., and Yirmiya, N. (1999). Mother-infant affect synchrony as an antecedent of the emergence of self-control. *Developmental psychology*, **35**:223–31. doi:10.1037//0012-1649.35.1.223.

Ferguson, T., Rowlands, A.V., Olds, T., and Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity*, **12**(1):42. doi:10.1186/s12966-015-0201-9.

Fisher, A.J. and Newman, M.G. (2013). Heart rate and autonomic response to stress after experimental induction of worry versus relaxation in healthy, high-worry, and generalized anxiety disorder individuals. *Biological psychology*, **93**(1):65–74.

Fitbit (n.d.). Fitbit one. accessed 29/01/2019.
URL `https://www.fitbit.com/no/one`

Fletcher, R.R., Poh, M.Z., and Eydgahi, H. (2010). Wearable sensors: Opportunities and challenges for low-cost health care. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 1763–1766. doi:10.1109/IEMBS.2010.5626734.

Fujiwara, K. (2016). Triadic synchrony: Application of multiple wavelet coherence to a small group conversation. *Applied Mathematics*, **7**:1477–1483. doi:10.4236/am.2016.714126.

Gabana, D., Tokarchuk, L., Hannon, E., and Gunes, H. (2017). Effects of valence and arousal on working memory performance in virtual reality gaming. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 36–41. doi:10.1109/ACII.2017.8273576.

Gao, W., Nyein, H.Y.Y., Shahpar, Z., Tai, L., Wu, E., Bariya, M., Ota, H., Fahad, H.M., Chen, K., and Javey, A. (2016). Wearable sweat biosensors. In *2016 IEEE International Electron Devices Meeting (IEDM)*. doi:10.1109/IEDM.2016.7838363.

Garbarino, M., Lai, M., Bender, D., Picard, R.W., and Tognetti, S. (2014). Empatica e3 - a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, pp. 39–42. doi:10.1109/MOBIHEALTH.2014.7015904.

Gashi, S., Di Lascio, E., and Santini, S. (2019). Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, **3**(1):13:1–13:19. doi:10.1145/3314400.

Gawley, R., Morrow, C., Chan, H., and Lindsay, R. (2016). Bitrun: Gamification of health data from fitbit® activity trackers. In M.U. Ahmed, S. Begum, and W. Raad (eds.), *Internet of Things Technologies for HealthCare*, pp. 77–82. Springer International Publishing.

Gianaros, P.J., Onyewuenyi, I.C., Sheu, L.K., Christie, I.C., and Critchley, H.D. (2012). Brain systems for baroreflex suppression during stress in humans. *Human brain mapping*, **33**(7):1700–1716.

Gillinov, S., Etiwy, M., Wang, R., Blackburn, G., Phelan, D., Gillinov, A.M., Houghtaling, P., Javadikasgari, H., and Desai, M.Y. (2017). Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine & Science in Sports & Exercise*, **49**(8):1697–1703. doi:10.1249/MSS.0000000000001284.

Gjoreski, M., Luštrek, M., Gams, M., and Gjoreski, H. (2017). Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, **73**:159 – 170. doi:https://doi.org/10.1016/j.jbi.2017.08.006.

Glauser, W. (2013). Doctors among early adopters of google glass. *CMAJ*, **185**(16):1385–1385. doi:10.1503/cmaj.109-4607.

Grafsgaard, J., Duran, N., Randall, A., Tao, C., and D'Mello, S. (2018). Generative multimodal models of nonverbal synchrony in close relationships. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 195–202. doi:10.1109/FG.2018.00037.

Greco, A., Valenza, G., Citi, L., and Scilingo, E.P. (2017). Arousal and valence recognition of affective sounds based on electrodermal activity. *IEEE Sensors Journal*, **17**(3):716–725. doi:10.1109/JSEN.2016.2623677.

Gribel, L., Regier, S., and Stengel, I. (2016). Acceptance factors of wearable computing: An empirical investigation. In *Proceedings of the Eleventh International Network Conference*, INC'16.

Guay, P., Gorgutsa, S., LaRochelle, S., and Messaddeq, Y. (2017). Wearable contactless respiration sensor based on multi-material fibers integrated into textile. *Sensors*, **17**(5). doi:10.3390/s17051050.

Haataja, E., Malmberg, J., and Järvelä, S. (n.d.). Monitoring in collaborative learning: Co-occurrence of observed behavior and physiological synchrony explored.

Haghi, M., Thurow, K., and Stoll, R. (2017). Wearable devices in medical internet of things: Scientific research and commercially available devices. *Healthcare Informatics Research*, **23**(1):4–15.

Hamilton-Craig, C., Fifoot, A., Hansen, M., Pincus, M., Chan, J., Walters, D.L., and Branch, K.R. (2014). Diagnostic performance and cost of ct angiography versus stress ecg – a randomized prospective study of suspected acute coronary syndrome chest pain in the emergency department (ct-compare). *International Journal of Cardiology*, **177**(3):867 – 873. doi:10.1016/j.ijcard.2014.10.090.

Han, D., Zhang, C., Fan, X., Hindle, A., Wong, K., and Stroulia, E. (2012). Understanding android fragmentation with topic analysis of vendor-specific bugs. In *Proceedings of the 2012 19th Working Conference on Reverse Engineering*, WCRE '12, pp. 83–92. IEEE Computer Society, Washington, DC, USA. doi:10.1109/WCRE.2012.18.

Hänsel, K., Alomainy, A., and Haddadi, H. (2016). Large scale mood and stress self-assessments on a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pp. 1180–1184. ACM, New York, NY, USA. doi:10.1145/2968219.2968305.

Hänsel, K., Haddadi, H., and Alomainy, A. (2017). Demo: Awsense: A framework for collecting sensing data from the apple watch. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '17, pp. 188–188. ACM, New York, NY, USA. doi:10.1145/3081333.3089333.

Hänsel, K., Katevas, K., Orgs, G., Richardson, D.C., Alomainy, A., and Haddadi, H. (2018a). The potential of wearable technology for monitoring social interactions based on interpersonal synchrony. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, pp. 45–47. ACM, New York, NY, USA. doi:10.1145/3211960.3211979.

Hänsel, K., Poguntke, R., Haddadi, H., Alomainy, A., and Schmidt, A. (2018b). What to put on the user: Sensing technologies for studies and physiology aware systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 145:1–145:14. ACM, New York, NY, USA. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173719.

Hänsel, K., Wilde, N., Haddadi, H., and Alomainy, A. (2015). Challenges with current wearable technology in monitoring health data and providing positive behavioural support. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, MOBIHEALTH'15, pp. 158–161. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium. ISBN 978-1-63190-088-4. doi:10.4108/eai.14-10-2015.2261601.

Hao, T., Walter, K.N., Ball, M.J., Chang, H.Y., Sun, S., and Zhu, X. (2017). Stresshacker: Towards practical stress monitoring in the wild with smartwatches. In *AMIA ... Annual Symposium proceedings. AMIA Symposium*, volume 2017, pp. 830–838.

Harrist, A.W. and Waugh, R.M. (2002). Dyadic synchrony: Its structure and function in children's development. *Developmental Review*, **22**(4):555–592. doi:10.1016/S0273-2297(02)00500-2.

Hassellund, S.S., Flaa, A., Sandvik, L., Kjeldsen, S.E., and Rostrup, M. (2010). Long-term stability of cardiovascular and catecholamine responses to stress tests an 18-year follow-up study. *Hypertension*, **55**(1):131–136.

Hassib, M., Khamis, M., Schneegass, S., Shirazi, A.S., and Alt, F. (2016). Investigating user needs for bio-sensing and affective wearables. In *Proceedings of the 2016 CHI Conference Extended*

Abstracts on Human Factors in Computing Systems, CHI EA '16, pp. 1415–1422. ACM, New York, NY, USA. ISBN 978-1-4503-4082-3. doi:10.1145/2851581.2892480.

Hayman, R., Taylor, B., Peart, N., Galland, B., and Sayers, R. (2001). Participation in research: Informed consent, motivation and influence. *Journal of Paediatrics and Child Health*, **37**. doi:10.1046/j.1440-1754.2001.00612.x.

Hernandez, J., Li, Y., Rehg, J., and Picard, R. (2014). Bioglass: Physiological parameter estimation using a head-mounted wearable device. In *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare*, MOBIHEALTH. doi:10.4108/icst.mobihealth.2014.257219.

Hernandez, J., McDuff, D., Infante, C., Maes, P., Quigley, K., and Picard, R. (2016). Wearable esm: Differences in the experience sampling method across wearable devices. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '16, pp. 195–205. ACM, New York, NY, USA. ISBN 978-1-4503-4408-1. doi:10.1145/2935334.2935340.

Hernandez, J., Morris, R.R., and Picard, R.W. (2011). Call center stress recognition with person-specific models. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 125–134. Springer.

Hiremath, S., Yang, G., and Mankodiya, K. (2014). Wearable internet of things: Concept, architectural components and promises for person-centered healthcare. In *Proceedings of the 2014 International Conference on Wireless Mobile Communication and Healthcare*, MobiHealth'14. doi:10.4108/icst.mobihealth.2014.257440.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., and Wood, K. (2006). Sensecam: A retrospective memory aid. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, pp. 177–193. doi:10.1007/11853565_11.

Hokanson, J.E. and Burgess, M. (1964). Effects of physiological arousal level, frustration, and task complexity on performance. *The Journal of Abnormal and Social Psychology*, **68**:698–702. doi:10.1037/h0049340.

Holler, J. (2010). Speakers' use of interactive gestures as markers of common ground. In S. Kopp and I. Wachsmuth (eds.), *Gesture in Embodied Communication and Human-Computer Interaction*, pp. 11–22.

Holt, M., Yule, B., Jackson, D., Zhu, M., and Moraveji, N. (2018). Ambulatory monitoring of respiratory effort using a clothing-adhered biosensor. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6.

Hooke, M., Gilchrist, L., Tanner, L., Hart, N., and Withycombe, J. (2016). Use of a fitness tracker to promote physical activity in children with acute lymphoblastic leukemia. *Journal of Pediatric Blood Cancer*, **63**(4). doi:10.1002/pbc.25860.

How, T.V., Chee, J., Wan, E., and Mihailidis, A. (2013). Mywalk: A mobile app for gait asymmetry rehabilitation in the community. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '13, pp. 73–76. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). doi:10.4108/icst.pervasivehealth.2013.252118.

Howcroft, J., Kofman, J., and Lemaire, E.D. (2017). Prospective fall-risk prediction models for older adults based on wearable sensors. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **25**(10):1812–1820. doi:10.1109/TNSRE.2017.2687100.

Huang, B., Li, M., Mei, T., McCoul, D., Qin, S., Zhao, Z., and Zhao, J. (2017). Wearable stretch sensors for motion measurement of the wrist joint based on dielectric elastomers. *Sensors*, **17**(12). doi:10.3390/s17122708.

Icahn School of Medicine at Mount Sinai (2015). Asthma health app | apple researchkit | icahn school of medicine at mount sinai. accessed 05/05/2015.
URL http://apps.icahn.mssm.edu/asthma/

IDC (2017). Forecast wearables unit shipments worldwide by product category from 2014 to 2021 (in millions). in statista - the statistics portal. accessed 22/11/2018.
URL https://www.statista.com/statistics/437879/wearables-worldwide-shipments-by-product-category/

Jiang, P. and Zhu, R. (2016). Dual tri-axis accelerometers for monitoring physiological parameters of human body in sleep. In *2016 IEEE SENSORS*, pp. 1–3. doi:10.1109/ICSENS.2016.7808735.

John Hopkins Medicine (2018). Johns hopkins epiwatch: App and research study. accessed 28/01/2019.
URL `https://www.hopkinsmedicine.org/epiwatch`

Julien, D., Brault, M., Chartrand, u., and Bégin, J. (2000). Immediacy behaviours and synchrony in satisfied and dissatisfied couples. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, **32**:84–90. doi:10.1037/h0087103.

Kamišalić, A., Fister, I., Turkanović, M., and Karakatič, S. (2018). Sensors and functionalities of non-invasive wrist-wearable devices: A review. *Sensors*, **18**(6). doi:10.3390/s18061714.

Karahanoğlu, A. and Erbuğ, c. (2011). Perceived qualities of smart wearables: Determinants of user acceptance. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces*, DPPI '11, pp. 26:1–26:8. ACM, New York, NY, USA. doi:10.1145/2347504.2347533.

Karapanos, E., Gouveia, R., Hassenzahl, M., and Forlizzi, J. (2016). Wellbeing in the making: Peoples' experiences with wearable activity trackers. *Psychology of Well-Being*, **6**(1):4. doi:10.1186/s13612-016-0042-6.

Karvonen, A., Kykyri, V.L., Kaartinen, J., Penttonen, M., and Seikkula, J. (2016). Sympathetic nervous system synchrony in couple therapy. *Journal of marital and family therapy*, **42 3**:383–95.

Kataoka, H., Kano, H., Yoshida, H., Saijo, A., Yasuda, M., and Osumi, M. (1998). Development of a skin temperature measuring system for non-contact stress evaluation.

Katevas, K., Haddadi, H., and Tokarchuk, L. (2014). Poster: Sensingkit – a multi-platform mobile sensing framework for large-scale experiments. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, pp. 375–378. doi:10.1145/2639108.2642910.

Katevas, K., Haddadi, H., Tokarchuk, L., and Clegg, R.G. (2015). Walking in sync: Two is company, three's a crowd. In *Proceedings of the 2Nd Workshop on Workshop on Physical Analytics*, WPA '15, pp. 25–29. ACM, New York, NY, USA. doi:10.1145/2753497.2753502.

Katevas, K., Hänsel, K., Clegg, R., Leontiadis, I., Haddadi, H., and Tokarchuk, L. (2018). Finding dory in the crowd: Detecting social interactions using multi-modal mobile sensing. *arXiv preprint arXiv:1809.00947*.

Kawsar, F., Min, C., Mathur, A., Van den Broeck, M., Acer, U.G., and Forlivesi, C. (2018). esense: Earable platform for human sensing. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '18, pp. 541–541. doi:10.1145/3210240.3211113.

Kelly, P., Marshall, S.J., Badland, H., Kerr, J., Oliver, M., Doherty, A.R., and Foster, C. (2013). An ethical framework for automated, wearable cameras in health behavior research. *American Journal of Preventive Medicine*, **44**(3):314 – 319. doi:https://doi.org/10.1016/j.amepre.2012.11.006.

Kendon, A. (1990). Conducting interaction: Patterns of behavior in focused encounters. *Studies in interactional sociolinguistics*, **7**.

King, M., D Burrows, G., and Stanley, G. (1983). Measurement of stress and arousal: Validation of the stress/arousal adjective checklist. *British journal of psychology (London, England : 1953)*, **74 (Pt 4)**:473–9. doi:10.1111/j.2044-8295.1983.tb01880.x.

Kirby, B., Kirby, A., and Birch, J.L. (2016). Wearable tech: Why architectures matter. In *Proceedings British HCI 2016 - Workshop on Human Centred Design for Intelligent Environments*, HCI '16.

Kjeldskov, J. and Skov, M.B. (2014). Was it worth the hassle?: Ten years of mobile hci research discussions on lab and field evaluations. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, MobileHCI '14, pp. 43–52. ACM. doi:10.1145/2628363.2628398.

Knight, J.F., Baber, C., Schwirtz, A., and Bristow, H.W. (2002). The comfort assessment of wearable computers. In *Proceedings. Sixth International Symposium on Wearable Computers*, volume 2, pp. 65–74.

Konvalinka, I., Xygalatas, D., Bulbulia, J., Schjødt, U., Jegindø, E.M., Wallot, S., Van Orden, G., and Roepstorff, A. (2011). Synchronized arousal between performers and related spectators in a fire-walking ritual. **108**(20):8514–8519. doi:10.1073/pnas.1016955108.

Kooiman, T.J.M., Dontje, M.L., Sprenger, S.R., Krijnen, W.P., van der Schans, C.P., and de Groot, M. (2015). Reliability and validity of ten consumer activity trackers. *BMC Sports Science, Medicine and Rehabilitation*, **7**(1):24. doi:10.1186/s13102-015-0018-5.
URL https://doi.org/10.1186/s13102-015-0018-5

Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pp. 279–328.

Kreitmair, K.V., Cho, M.K., and Magnus, D.C. (2017). Consent and engagement, security, and authentic living using wearable and mobile health technology. *Nature Biotechnology*, **35**(7):617–620. doi:10.1038/nbt.3887.

Lake, S. (2018). Ending sales of myo, preparing for the future. accessed 03/03/2019.
URL https://medium.com/@srlake/ending-sales-of-myo-preparing-for-the-future-281af9bbcac2

Larson, R. and Csikszentmihalyi, M. (1983). The experience sampling method. In H.T. Reis (ed.), *New Directions for Methodology of Social and Behavioural Sciences*, pp. 41–56.

Launay, J., Tarr, B., and Dunbar, R.I.M. (n.d.). Synchrony as an adaptive mechanism for large-scale human social bonding. *Ethology*, **122**(10):779–789. doi:10.1111/eth.12528.

Ledger, D. and McCaffrey, D. (2014). Inside wearables: How the science of human behavior change offers the secret to long-term engagement. Online posted January-2014.
URL https://endeavour.partners/white-papers/

LEE, J., CHO, S., LEE, J., LEE, K., and YANG, H. (2007). Wearable accelerometer system for measuring the temporal parameters of gait. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 483–486. doi:10.1109/IEMBS.2007.4352328.

LEE, S.A., SPENCER, D.D., and SPENCER, S.S. (2000). Intracranial eeg seizure-onset patterns in neocortical epilepsy. *Epilepsia*, **41**(3):297–307. doi:10.1111/j.1528-1157.2000.tb00159.x.

LEPINE, N.N., TAJIMA, T., OGASAWARA, T., KASAHARA, R., and KOIZUMI, H. (2016). Robust respiration rate estimation using adaptive kalman filtering with textile ecg sensor and accelerometer. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3797–3800. doi:10.1109/EMBC.2016.7591555.

LIKERT, R. (1932). A technique for the measurement of attitudes. In R.S. Woodworth (ed.), *Archives of Psychology*, volume 22, pp. 5–55.

LINDEN, W. (1991). What do arithmetic stress tests measure? protocol variations and cardiovascular responses. *Psychophysiology*, **28**(1):91–102.

LOEWEN, P.J., LYLE, G., and NACHSHEN, J.S. (2010). *An Eight-Item Form of the Empathy Quotient (EQ) and an Application to Charitable Giving*.
URL `http://individual.utoronto.ca/loewen/Research_files/Eight%20Question%20ES_final.pdf`

LUBANS, D.R., MORGAN, P.J., COLLINS, C.E., BOREHAM, C.A., and CALLISTER, R. (2009). The relationship between heart rate intensity and pedometer step counts in adolescents. *Journal of Sports Sciences*, **27**(6):591–597. doi:10.1080/02640410802676687.

LUNKENHEIMER, E., TIBERIO, S.S., SKORANSKI, A.M., BUSS, K.A., and COLE, P.M. (2018). Parent-child coregulation of parasympathetic processes varies by social context and risk for psychopathology. *Psychophysiology*, **55**(2). doi:10.1111/psyp.12985.

LYKKEN, D.T., ROSE, R., LUTHER, B., and MALEY, M. (1966). Correcting psychophysiological measures for individual differences in range. *Psychological Bulletin*, **66**:481–484. doi:10.1037/h0023922.

MacLean, D.L., Roseway, A., and Czerwinski, M. (2013). Moodwings: a wearable biofeedback device for real-time stress intervention. *PETRA*, pp. 66–68. doi:10.1145/2504335.2504406.

Mair, A., Poirier, M., and Conway, M.A. (2018). Author accepted manuscript: Memory for staged events: supporting older and younger adults' memory with sensecam. *Quarterly Journal of Experimental Psychology*. doi:10.1177/1747021818765038. Mair, A., Poirier, M. & Conway, M. A., Memory for staged events: supporting older and younger adults' memory with SenseCam, Quarterly Journal of Experimental Psychology. Copyright © 2018, the authors. Reprinted by permission of SAGE Publications.

Malmivaara, M. (2009). The emergence of wearable computing. In J. McCann and D. Bryson (eds.), *Smart Clothes and Wearable Technology*, Woodhead Publishing Series in Textiles, pp. 3 – 24. Woodhead Publishing.

Mandal, B., Chia, S.C., Li, L., Chandrasekhar, V., Tan, C., and Lim, J.H. (2015). A wearable face recognition system on google glass for assisting social interactions. In C.V. Jawahar and S. Shan (eds.), *Computer Vision - ACCV 2014 Workshops*, pp. 419–433. Springer International Publishing, Cham.

Mann, H. and Whitney, D. (1947). Individual comparisons by ranking methods. *The Annals of Mathematical Statistics*, **18**:50–60.

Marci, C.D., Ham, J., Moran, E., and Orr, S.P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of Nervous and Mental Disease*, **195**(2):103–111. doi:10.1097/01.nmd.0000253731.71025.fc.

McDuff, D.J., Hernandez, J., Gontarek, S., and Picard, R.W. (2016). Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4000–4004. ACM.

Mehrabian, A. and Russell, J.A. (1974). *An Approach to Environmental Psychology*. ISBN 0262130904.

de Mendonça, J.S., Cossette, L., Strayer, F.F., and Gravel, F. (2010). Mother-child and father-child interactional synchrony in dyadic and triadic interactions. *Sex Roles*, **64**(1-2):132–142. doi:10.1007/s11199-010-9875-2.

Mercer, K., Li, M., Giangregorio, L., Burns, C., and Grindrod, K. (2016). Behavior change techniques present in wearable activity trackers: A critical analysis. *JMIR mHealth uHealth 2016*, **4**(2). doi:10.2196/mhealth.4461.

Merletti, R., Parker, P., and Parker, P. (2004). *Electromyography: Physiology, Engineering, and Non-Invasive Applications*. IEEE Press Series on Biomedical Engineering. Wiley.

Microsoft (2016). Microsoft Band SDK - SDK Documentation. accessed 13/10/2018.
URL `https://developer.microsoftband.com/Content/docs/Microsoft%20Band%20SDK.pdf`

Mind Media (n.d.). Nexus-10 mkii. Last accessed: 08/10/2018.
URL `https://www.mindmedia.com/en/products/nexus-10-mkii/`

Mind Media (2017). Nexus-10 mkii specifications. accessed 15/12/2017.
URL `https://www.mindmedia.com/products/nexus-10-mkii/`

MindMedia (2017). User Manual - BioTrace+ Software. accessed 13/10/2018.
URL `http://81.169.139.131/download/documentation/UserManual_MM_BioTrace-V2017A_EN.pdf`

Mitkidis, P., McGraw, J.J., Roepstorff, A., and Wallot, S. (2015). Building trust: Heart rate synchrony and arousal during joint action increased by public goods game. *Physiology & Behavior*, **149**:101 – 106. doi:10.1016/j.physbeh.2015.05.033.

Mok, T.M., Cornish, F., and Tarr, J. (2015). Too much information: Visual research ethics in the age of wearable cameras. *Integrative Psychological and Behavioral Science*, **49**(2):309–322. doi:10.1007/s12124-014-9289-8.

Montanari, A., Tian, Z., Francu, E., Lucas, B., Jones, B., Zhou, X., and Mascolo, C. (2018). Measuring interaction proxemics with wearable light tags. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, **2**(1):25:1–25:30. doi:10.1145/3191757.

Moore, G.E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, **86**(1):82–85. doi:10.1109/JPROC.1998.658762.

Morgan, E., Gunes, H., and Bryan-Kinns, N. (2015). Using affective and behavioural sensors to explore aspects of collaborative music making. *International Journal of Human-Computer Studies*, **82**:31 – 47. doi:10.1016/j.ijhcs.2015.05.002.

Morris, J.D. (1995). Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, **35**(6):63–68.

Mozos, O.M., Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Dobrescu, R., and Manuel, J. (2017). Stress detection using wearable physiological and sociometric sensors. *International Journal of Neural Systems*, **27**(02). doi:10.1142/S0129065716500416.

Murakami, H., Kawakami, R., Nakae, S., Nakata, Y., Ishikawa-Takata, K., Tanaka, S., and Miyachi, M. (2016). Accuracy of wearable devices for estimating total energy expenditure: Comparison with metabolic chamber and doubly labeled water methodaccuracy of wearable devices for estimating total energy expenditureletters. *JAMA Internal Medicine*, **176**(5):702–703. doi:10.1001/jamainternmed.2016.0152.

Nagaoka, C. and Komori, M. (2008). Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. *IEICE Transactions on Information and Systems*, **E91-D**(6):1634–1640. doi:10.1093/ietisy/e91-d.6.1634.

Nagaoka, C., Komori, M., and Yoshikawa, S. (2005). Synchrony tendency: Interactional synchrony and congruence of nonverbal behavior in social interaction. In *International Conference on Active Media Technology*, AMT 2005, pp. 529–534. ISBN 0-7803-9035-0. doi:10.1109/AMT.2005.1505415.

Najström, M. and Jansson, B. (2007). Skin conductance responses as predictor of emotional responses to stressful life events. *Behaviour Research and Therapy*, **45**(10):2456 – 2463. doi:10.1016/j.brat.2007.03.001.

Natarajan, A., Xu, K.S., and Eriksson, B. (2016). Detecting divisions of the autonomic nervous system using wearables. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5761–5764. doi:10.1109/EMBC.2016.7592036.

Nerino, R., Contin, L., da Silva Pinto, W.J.G., Massazza, G., Actis, M., Capacchione, P., Chimienti, A., and Pettiti, G. (2013). A bsn based service for post-surgical knee rehabilitation at home. In *Proceedings of the 8th International Conference on Body Area Networks*, BodyNets '13, pp. 401–407. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). doi:10.4108/icst.bodynets.2013.253679.

Nielsen, J. (1994). Ten usability heuristics.
URL https://www.nngroup.com/articles/ten-usability-heuristics/

Nintendo (n.d.). Pokémon systems. accessed 29/01/2019.
URL https://www.nintendo.com/consumer/systems/other/pokemon.jsp

NYU Langone Health (2018). Concussion tracker app. accessed 28/01/2019.
URL https://nyulangone.org/apps/concussion-tracker-app

Packer, H.S., Buzogany, G., Smith, D.A., Dragan, L., Van Kleek, M., and Shadbolt, N.R. (2014). The editable self: A workbench for personal activity data. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pp. 2185–2190. ACM, New York, NY, USA. ISBN 978-1-4503-2474-8. doi:10.1145/2559206.2581283.

Palumbo, R.V., Marraccini, M.E., Weyandt, L.L., Wilder-Smith, O., McGee, H.A., Liu, S., and Goodwin, M.S. (2016). Interpersonal autonomic physiology: a systematic review of the literature. *Personality and Social Psychology Review*, **21**(2):99–141. doi:10.1177/1088868316628405.

Pantelopoulos, A. and Bourbakis, N.G. (2010). A survey on wearable sensor-based systems for health monitoring and prognosis. *Trans. Sys. Man Cyber Part C*, **40**(1):1–12. doi:10.1109/TSMCC.2009.2032660.

Park, T., Lee, U., Lee, B., Lee, H., Son, S., Song, S., and Song, J. (2013). Exersync: Facilitating interpersonal synchrony in social exergames. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pp. 409–422. doi:10.1145/2441776.2441823.

Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, **9**(1):21. doi:10.1186/1743-0003-9-21.

Peake, J.M., Kerr, G., and Sullivan, J.P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology*, **9**:743. doi:10.3389/fphys.2018.00743.

Pentland, A.P. (1998). Wearable intelligence. *Scientific American*, **9**(4):90–95.

Pierleoni, P., Belli, A., Palma, L., Pellegrini, M., Pernini, L., and Valenti, S. (2015). A high reliability wearable device for elderly fall detection. *IEEE Sensors Journal*, **15**(8):4544–4553. doi:10.1109/JSEN.2015.2423562.

Pietilä, J., Mehrang, S., Tolonen, J., Helander, E., Jimison, H., Pavel, M., and Korhonen, I. (2017). *Evaluation of the Accuracy and Reliability for Photoplethysmography Based Heart Rate and Beat-to-Beat Detection During Daily ActivitiesIFMBE Proceedings*, pp. 145–148. ISBN 978-981-10-5121-0. doi:10.1007/978-981-10-5122-7_37.

Pogunktke, R., Hänsel, K., Haddadi, H., Alomainy, A., and Schmidt, A. (2018). Developing the design space for physiological measurement tools: A theoretical foundation of decision criteria. *Under review: International Journal of Human-Computer Studies*.

Poh, M.Z., Loddenkemper, T., Reinsberger, C., Swenson, N.C., Goyal, S., Sabtala, M.C., Madsen, J.R., and Picard, R.W. (2012). Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia*, **53**(5):e93–e97. doi:10.1111/j.1528-1167.2012.03444.x.

Poh, M.Z., Swenson, N.C., and Picard, R.W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE transactions on bio-medical engineering*, **57**(5):1243–1252. doi:10.1109/TBME.2009.2038487.

Polar (n.d.). Polar h7 user manual. accessed 02/02/2019.
URL https://support.polar.com/e_manuals/H7_Heart_Rate_Sensor/Polar_H7_Heart_Rate_Sensor_accessory_manual_English__.pdf

Polar (2018). H6, H7, H10 and OH1 Heart rate sensors. accessed 13/10/2018.
URL http://developer.polar.com/wiki/H6,_H7,_H10_and_OH1_Heart_rate_sensors

Polar Support (2017). How to check the battery level status of my heart rate sensor? accessed 15/12/2017.
URL https://support.polar.com/en/support/how_to_check_the_battery_level_status_of_my_heart_rate_sensor

Profita, H., Clawson, J., Gilliland, S.M., Zeagler, C., Starner, T., Budd, J., and Do, E.Y.L. (2013). Don't mind me touching my wrist: a case study of interacting with on-body technology in public. *ISWC*, p. 89. doi:10.1145/2493988.2494331.

Quek, M., Boland, D., Williamson, J., Murray-Smith, R., Tavella, M., Perdikis, S., Schreuder, M., and Tangermann, M. (2011). Simulating the feel of brain-computer interfaces for design, development and social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 25–28. doi:10.1145/1978942.1978947.

Quesada, A.A., Tristao, R.M., Pratesi, R., and Wolf, O.T. (2014). Hyper-responsiveness to acute stress, emotional problems and poorer memory in former preterm children. *Stress*, **17**(5):389–399.

Quiroz, J.C., Geangu, E., and Yong, M.H. (2018). Emotion-recognition using smart watch sensor data: Mixed-design study. *JMIR mental health*, **5**(3):1–17.

Ragot, M., Martin, N., Em, S., Pallamin, N., and Diverrez, J.M. (2017). Emotion recognition using physiological signals: Laboratory vs. wearable sensors. *International Conference on Applied Human Factors and Ergonomics*, pp. 15–22.

Rammstedt, B. and John, O.P. (2007). Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, **41**(1):203–212. doi:10.1016/j.jrp.2006.02.001.

Ramseyer, F. and Tschacher, W. (2014). Nonverbal synchrony of head- and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in Psychology*, **5**:979. doi:10.3389/fpsyg.2014.00979.

Reddish, P., Fischer, R., and Bulbulia, J. (2013). Let's dance together: Synchrony, shared intentionality and cooperation. *PLOS ONE*, **8**(8):1–13. doi:10.1371/journal.pone.0071182. URL `10.1371/journal.pone.0071182`

Redmond, S.J., Lovell, N.H., Yang, G.Z., Horsch, A., Lukowicz, P., Murrugarra, L., and Marschollek, M. (2014). What does big data mean for wearable sensor systems? contribution of the imia wearable sensors in healthcare. *Yearbook of medical informatics*, **9**:135–42.

Regal, G., Wais-Zechmann, B., Gattol, V., Garschall, M., and Tscheligi, M. (2018). Smart pocket watch: Exploring the design space for wearable technology in healthcare. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference on*, pp. 269–272.

Rogers, Y. and Marshall, P. (2017). Research in the wild. *Synthesis Lectures on Human-Centered Informatics*. doi:10.2200/S00764ED1V01Y201703HCI03.

Rosenberger, M.E., Buman, M.P., Haskell, W.L., McConnell, M.V., and Carstensen, L.L. (2016). 24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices. *Journal of Med Sci Sports Exerc*, **48**(3):457 – 465. doi:10.1249/MSS.0000000000000778.

Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *J. Information Science*, **33**:163–180.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, **39**(6):1161–1178.

Sakakibara, R., Terayama, K., Ogawa, A., Haruta, H., Akiba, T., Tateno, F., Kishi, M., Tsuyusaki, Y., Aiba, Y., and Ogata, T. (2017). Wearable gait sensors to measure degenerative cerebellar ataxia. *Journal of the Neurological Sciences*, **381**:56 – 57. doi:10.1016/j.jns.2017.08.219. Abstracts from the World Congress of Neurology (WCN 2017).

Salimpoor, V.N., Benovoy, M., Longo, G., Cooperstock, J.R., and Zatorre, R.J. (2009). The rewarding aspects of music listening are related to degree of emotional arousal. *PloS one*, **4**(10):e7487.

Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., and Mozos, O. (2015). Stress detection using wearable physiological sensors. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pp. 526–532. ISBN 978-3-319-18913-0. doi:10.1007/978-3-319-18914-7_55.

Sano, A. and Picard, R.W. (2013). Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 671–676. doi:10.1109/ACII.2013.117.

Santos, J. (2016). The rights and rightness of mhealth data. accessed 10/02/2019.
URL `https://us.kantar.com/business/health/2016/mhealth-data-ownership-and-accuracy/`

Schaefer, H.S., Larson, C.L., Davidson, R.J., and Coan, J.A. (2014). Brain, body, and cognition: Neural, physiological and self-report correlates of phobic and normative fear. *Biological psychology*, **98**:59–69.

Schimmack, U. and Grob, A. (2000). Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. *European Journal of Personality*, **14**(4):325–345. doi:10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.3.CO;2-9.

Schneiderman, N., Ironson, G., and Siegel, S.D. (2005). Stress and health: psychological, behavioral, and biological determinants. *Annu. Rev. Clin. Psychol.*, **1**:607–628.

Segura Anaya, L.H., Alsadoon, A., Costadopoulos, N., and Prasad, P.W.C. (2018). Ethical implications of user perceptions of wearable devices. *Science and Engineering Ethics*, **24**(1):1–28. doi:10.1007/s11948-017-9872-8.

Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., Hassan, M., and Seneviratne, A. (2017). A survey of wearable devices and challenges. *IEEE Communications Surveys Tutorials*, **19**(4):2573–2620. doi:10.1109/COMST.2017.2731979.

Seraganian, P., Szabo, A., and Brown, T.G. (1997). The effect of vocalization on the heart rate response to mental arithmetic. *Physiology & behavior*, **62**(2):221–224.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on information technology in biomedicine*, **14**(2):410–417.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**:591–611.

Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J.G., Vhaduri, S., Preston, K., and Kumar, S. (2015). Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pp. 505–516. ACM, New York, NY, USA. ISBN 978-1-4503-3574-4. doi:10.1145/2750858.2807537.

Shcherbina, A., Mattsson, C.M., and Waggott, D. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, **7**(2):3. doi:10.3390/jpm7020003.

Shelgikar, A.V., Anderson, P.F., and Stephens, M.R. (2016). Sleep tracking, wearable technology, and opportunities for research and clinical care. *Chest*, **150**(3):732 – 743. doi:10.1016/j.chest.2016.04.016.

Shibuya, N., Nukala, B.T., Rodriguez, A.I., Tsay, J., Nguyen, T.Q., Zupancic, S., and Lie, D.Y.C. (2015). A real-time fall detection system using a wearable gait analysis sensor and a support vector machine (svm) classifier. In *2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*, pp. 66–67. doi:10.1109/ICMU.2015.7061032.

Slovák, P., Tennent, P., Reeves, S., and Fitzpatrick, G. (2014). Exploring skin conductance synchronisation in everyday interactions. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, pp. 511–520. doi:10.1145/2639189.2639206.

Spierer, D.K., Rosen, Z., and Litman, L.L. (2015). Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of Medical Engineering & Technology*, **39**(5):264–271. doi:10.3109/03091902.2015.1047536.

Stahl, S.E., An, H.S., Dinkel, D.M., Noble, J.M., and Lee, J.M. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? are they accurate enough? *BMJ Open Sport & Exercise Medicine*, **2**(1):e000106. doi:10.1136/bmjsem-2015-000106.

Stanford Medicine (2018). Apple heart study. accessed 28/01/2019.
URL `https://med.stanford.edu/appleheartstudy.html`

Statista (n.d.). Wearable technology - statistics & facts. Retrieved:07/18/2018.
URL `https://www.statista.com/topics/1556/wearable-technology/`

Suveg, C., Shaffer, A., and Davis, M. (2016). Family stress moderates relations between physiological and behavioral synchrony and child self-regulation in mother–preschooler dyads.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, **1**(2):58–99. doi:10.1089/big.2012.0002.

Swearingen, J. (2017). The rebirth of google glass on the factory floor. accessed 04/02/2019.
URL `http://nymag.com/intelligencer/2017/07/the-rebirth-of-google-glass-on-the-factory-floor.html`

Tamura, T., Maeda, Y., Sekine, M., and Yoshida, M. (2014). Wearable photoplethysmographic sensors – past and present. *Electronics*, **3**(2):282–302. doi:10.3390/electronics3020282.

Tarr, B., Launay, J., Cohen, E., and Dunbar, R. (2015). Synchrony and exertion during dance independently raise pain threshold and encourage social bonding. *Biology Letters*, **11**(10):20150767. doi:10.1098/rsbl.2015.0767.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Technical report, Lippincott Williams & Wilkins. doi:10.1161/01.CIR.93.5.1043.

Tech Crunch (2015). Tim cook says apple watch ships in april. accessed 22/01/2019. URL https://techcrunch.com/2015/01/27/tim-cook-says-apple-watch-should-ship-in-april/

Thought Technology Ltd. (2010). Basics of heart rate variability applied to psychophysiology. Technical report.

Tomaka, J., Blascovich, J., and Swart, L. (1994). Effects of vocalization on cardiovascular and electrodermal responses during mental arithmetic. *International Journal of Psychophysiology*, **18**(1):23–33.

Torres, C.A., Orozco, A.A., and Álvarez, M.A. (2013). Feature selection for multimodal emotion recognition in the arousal-valence space. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4330–4333. doi:10.1109/EMBC.2013.6610504.

Tourunen, A. (2017). *Sympathetic nervous system synchrony between participants of couple therapy*. Ph.D. thesis.

Tunçgenç, B. and Cohen, E. (2016). Movement synchrony forges social bonds across group divides. *Frontiers in Psychology*, **7**. doi:10.3389/fpsyg.2016.00782.

Vacharkulksemsuk, T. and Fredrickson, B.L. (2012). Strangers in sync: Achieving embodied rapport through shared movements. *Journal of Experimental Social Psychology*, **48**(1):399–402. doi:10.1016/j.jesp.2011.07.015.

Valdesolo, P., Ouyang, J., and DeSteno, D. (2010). The rhythm of joint action: Synchrony promotes cooperative ability. *Journal of Experimental Social Psychology*, **46**(4):693–695. doi:10.1016/j.jesp.2010.03.004.

Vanutelli, M.E., Gatt, L., Angioletti, L., , and Balconi, M. (2017). Affective synchrony and autonomic coupling during cooperation: A hyperscanning study. *BioMed Research International*, **2017**. doi:10.1155/2017/3104564.

Vinkers, C.H., Penning, R., Hellhammer, J., Verster, J.C., Klaessens, J.H., Olivier, B., and Kalkman, C.J. (2013). The effect of stress on core and peripheral body temperature in humans. *Stress*, **16**(5):520–530.

Vlemincx, E., Van Diest, I., and Van den Bergh, O. (2012). A sigh following sustained attention and mental stress: Effects on respiratory variability. *Physiology & Behavior*, **107**(1):1–6. doi:10.1016/j.physbeh.2012.05.013.

Wallen, M.P., Gomersall, S.R., Keating, S.E., Wisløff, U., and Coombes, J.S. (2016). Accuracy of heart rate watches: Implications for weight management. *PLoS ONE*, **11**(5):e0154420. doi:10.1371/journal.pone.0154420.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A.T. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pp. 3–14. ACM Press. doi:10.1145/2632048.2632054.

Wang, W., Bryan-kinns, N., and Yan, Q. (2015). The design space and the shifting trigger in wearable product development. In *International Design Conference of KSDS and ADADA with Cumulus*, IDC'15, pp. 206–210.

Ward, J.A., Richardson, D., Orgs, G., Hunter, K., and Hamilton, A. (2018). Sensing interpersonal synchrony between actors and autistic children in theatre using wrist-worn accelerometers. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ISWC '18, pp. 148–155. ACM, New York, NY, USA. doi:10.1145/3267242.3267263.

Warren, T. (2019). Microsoft to offer band refunds, announces end of apps and services. accessed 07/05/2019.
URL `https://www.theverge.com/2019/3/1/18246794/microsoft-band-refunds-apps-services-end-of-support-may-fitness-tracker`

Waters, S., V. West, T., and Mendes, W. (2014). Stress contagion. *Psychological Science*. doi:10.1177/0956797613518352.

Watson, D., Clark, L.A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scale. *Journal of Personality and Social Psychology*, **54**:1063–1070.

Watson, D., Wiese, D., Vaidya, J., and Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of personality and social psychology*, **76**(5):820–838. doi:10.1037//0022-3514.76.5.820.

Weitkunat, R., Coggins, C.R., Sponsiello-Wang, Z., Kallischnigg, G., and Dempsey, R. (2013). Assessment of cigarette smoking in epidemiologic studies. *Beiträge zur Tabakforschung/Contributions to Tobacco Research*, **25**(7):638–648.

Wijkstra, P.J., Ten Vergert, E.M., van Altena, R., Otten, V., Kraan, J., Postma, D.S., and Koëter, G.H. (1995). Long term benefits of rehabilitation at home on quality of life and exercise tolerance in patients with chronic obstructive pulmonary disease. *Thorax*, **50**(8):824–828. doi:10.1136/thx.50.8.824.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, **1**(6):80–83.

Wired (2018). Apple watch 4 adds ecg, ekg, and more heart-monitoring capabilities. accessed 22/01/2019.
URL https://www.wired.com/story/apple-watch-series-4/

Witvliet, C.V. and Vrana, S.R. (2007). Play it again sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial emg, and heart rate. *Cognition and Emotion*, **21**(1):3–25.

Won, A.S., Bailenson, J.N., Stathatos, S.C., and Dai, W. (2014). Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *Journal of Nonverbal Behavior*, **38**(3):389–408.

Wood, L.B. and Asada, H.H. (2007). Low variance adaptive filter for cancelling motion artifact in wearable photoplethysmogram sensor signals. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 652–655. doi:10.1109/IEMBS.2007.4352374.

Yamamoto, A., Nakamoto, H., Yamaji, T., Ootaka, H., Bessho, Y., Nakamura, R., and Ono, R. (2017). Method for measuring tri-axial lumbar motion angles using wearable sheet stretch sensors. *PLOS ONE*, **12**(10):1–14. doi:10.1371/journal.pone.0183651.

Yang, J., Wang, S., Chen, N., Chen, X., and Shi, P. (2010). Wearable accelerometer based extendable activity recognition system. In *2010 IEEE International Conference on Robotics and Automation*, pp. 3641–3647. doi:10.1109/ROBOT.2010.5509783.

Zhu, Y. (2015). Researchkit - glucosuccess. accessed 15/11/2015.
URL `https://github.com/ResearchKit/GlucoSuccess`

Zhu, Z., Satizábal, H.F., Blanke, U., Perez-Uribe, A., and Tröster, G. (2016). Naturalistic recognition of activities and mood using wearable electronics. *IEEE Transactions on Affective Computing*, **7**(3):272–285. doi:10.1109/TAFFC.2015.2491927.

von Zimmermann, J., Vicary, S., Sperling, M., Orgs, G., and Richardson, D.C. (2018). The choreography of group affiliation. *Topics in Cognitive Science*, **10**(1):80–94. doi:10.1111/tops.12320.